





UNIVERSIDAD DE SALAMANCA, UNIVERSIDAD DE SANTIAGO, UNIVERSIDAD DE LA CORUÑA, UNIVERSIDAD DE LA LAGUNA, UNIVERSIDAD DE VALLADOLID, UNIVERSIDAD DE VALENCIA, CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS.

ANÁLISIS DE UN AUTÓMATA CLASIFICADOR DE IMÁGENES.

IMPLICACIONES FILOSÓFICAS SOBRE LA ATRIBUCIÓN DE CONCEPTOS.

Tesis doctoral presentada por Juan Miguel Benavent Gomar

dentro del Programa de Doctorado en Lógica y Filosofía de la Ciencia

Dirigida por el Dr. Valeriano Iranzo García

València, Febrero de 2017

Universidad de Salamanca, Universidad de Santiago, Universidad de La Coruña, Universidad de La Laguna, Universidad de Valladolid, Universidad de Valencia, Consejo Superior de Investigaciones Científicas.

Autor: Juan Miguel Benavent Gomar

Director: Dr. Valeriano Iranzo García

Texto impreso en València

Primera edición,

A la memoria de mi padre,
auténtico ejemplo de espíritu libre.

Resumen

El presente trabajo se ocupa de la construcción de un clasificador visual automático (un programa de ordenador, en realidad) de imágenes basado en regresiones logísticas. Se pretende, además, comprobar su eficacia en la realización de dicha tarea (la clasificación de imágenes), y reflexionar sobre las implicaciones filosóficas de los resultados obtenidos. En la introducción apuntaré, en términos generales, la problemática que suscita el reconocimiento de imágenes por ordenador. Después se plantea una situación concreta, en un escenario típico de recuperación de información visual, donde se realiza un experimento que nos permitirá extraer datos empíricos (capítulo 2), y a continuación describiremos el marco teórico actual acerca de lo que se entiende por concepto (capítulos 3 y 4). En principio se trata de comprobar hasta qué punto un ordenador es capaz de reconocer imágenes, aunque tal vez se podría ir más allá y decir que lo que realmente está en juego aquí es la capacidad de un ordenador para aprender conceptos. Por eso en el capítulo 5 se discuten las implicaciones filosóficas, esto es, sobre la naturaleza y la adquisición de los conceptos, de los resultados experimentales.

Palabras clave: Inteligencia artificial, concepto, categorización, representación, regresiones logísticas, neo-empirismo.

Agradecimientos

"Si he visto más lejos es porque estoy sentado

sobre los hombros de gigantes.

Bernardo de Chartres (s.XII),
reutilizada por Isaac Newton.

En primer lugar, he de agradecer los consejos de mi director de tesis, el Dr. Valeriano Iranzo García, cuya ayuda y apoyo han sido cruciales para la consecución de esta tesis.

A mis compañeros de equipo de investigación en inteligencia Artificial, UNED-UV, con quienes me he formado en las tareas de investigación, en especial a la Dra. Xaro Benavent García, con quien mantengo lazos telúricos y de amistad desde hace muchísimo tiempo, y que fue la que me impulsó a iniciar las investigaciones en el campo de la recuperación de imágenes.

A mis compañeros de investigación de lógica y filosofía de la ciencia del grupo Méthodos, por sus aportaciones críticas y por haber compartido la pasión por esta disciplina.

A mis compañeros informáticos, Dr. Miguel Arevarillo y Luís Marco, por tantos días de discusiones enriquecedoras detrás de buenas comidas y buen vino.

Finalmente a la paciencia y tolerancia inagotable de mi familia, verdadero motor de todas mis victorias.

A todos ellos, y a todos los que no he nombrado pero que realmente deberían aparecer si las listas no tuvieran la restricción de finitud,

gracias,

Joan Benavent

Índice general

Índice de figuras

Índice de cuadros

Hay tantas realidades como puntos de vista. El punto de vista crea el panorama.

José Ortega y Gasset

CAPÍTULO

1

Introducción: la clasificación de imágenes

A medida que las bases de datos de imágenes han ido creciendo se han acentuado los problemas de recuperación de estas, es decir, cómo extraer la imagen deseada en una base de datos. Los modelos típicos en los que una etiqueta y una imagen estaban conectadas como una aplicación biyectiva solo son útiles si se conoce la etiqueta. Además, cuando un usuario busca una imagen, en general, no debería tener la obligación de saber ni la etiqueta asignada por la base de datos, ni mucho menos la imagen que desea. Digamos que busca algo acerca de perros, y pretende que el sistema muestre –recupere– las imágenes relacionadas con este concepto. ¿Se deberían reetiquetar las imágenes con estos conceptos, aparte de su etiqueta original? Si el solicitante es un humano, dotado de una ontología de conceptos, deseará que la clasificación intrínseca de la base de datos responda a su categorización ontológica. Pero si el concepto que busca no es tan simple como 'perro', o sencillamente, la base de datos no lo aplica como principio de clasificación, el problema de la recuperación se complica.

En este sentido, las grandes empresas suministradoras de imágenes suelen contratar mano de obra barata para etiquetarlas, siendo el criterio del humano etiquetador el que resuelve la etiqueta conceptual de cada imagen. A este procedimiento

le podemos hacer, al menos, dos objeciones: (i) ¿son universales los conceptos etiquetados?, es decir, ¿poseen el mismo contenido, o se interpretan igual, por parte de todos los humanos? (ii) ¿está libre de errores el etiquetador humano?

Sobre la primera objeción, y aun suponiendo que los organismos internacionales de estándares consiguieran ponerse de acuerdo en realizar una ontología normativa común, cabe apuntar que la velocidad de crecimiento conceptual del ser humano haría obsoleta tal ontología; además, con la lentitud burocrática con que suelen trabajar estos organismos, las ontologías, o sus modificaciones, nacerían desfasadas. Todavía más. Suponiendo que se superaran satisfactoriamente los problemas planteados por la estandarización de las ontologías, ¿se deberían reetiquetar todas las bases de datos con cada modificación de la ontología o debería el sujeto humano ajustarse a la categorización ontológica subyacente en la base de datos? Aun así, ¿y si la búsqueda se realiza en diferentes bases de datos con ontologías diferentes?

Respecto a la segunda objeción, según la evidencia experimental aportada por la psicología cognitiva, y cualquiera que sea la teoría de conceptos utilizada para enmarcar esos experimentos, hemos de afrontar el problema de la ignorancia y el error. En una base de datos que contiene una cantidad ingente de imágenes, el sujeto puede encontrase con algunas que no sabe a qué concepto vincular. Y salvo que demos por buena la infalibilidad del sujeto, algo poco realista si se tienen en cuenta las condiciones en las que se realizan estas tareas, habremos de admitir la posibilidad de errores debidos a múltiples causas. Y aún queda otro factor relevante que puede influir en los resultados. Piénsese en imágenes que diferentes sujetos, en función de sus condiciones socioculturales, sus ideologías o creencias religiosas, etc., adscribirían a conceptos diferentes, con lo cual la clasificación resultante no estará exenta de cierto componente subjetivo y no excluye divergencias entre sujetos a propósito de la misma base de imágenes y la etiqueta conceptual en cuestión.

Por todo lo expuesto anteriormente, entonces, no parece que las etiquetas resuelvan el problema, aunque hay mucho trabajo realizado en este sentido, sin más que ver cómo funciona el clasificador de imágenes de Google.

Una dificultad añadida es que no todos los conceptos pueden expresarse por medio de etiquetas simples. Basta formular una búsqueda un tanto compleja para entender la dificultad de este planteamiento: “Quiero encontrar una palmera coco-

tera en una isla paradisíaca, al atardecer en un día nublado. Necesito esta imagen para colgarla en mi habitación". En fin, ¿qué otras estrategias podrían ser útiles, además del etiquetado textual, en casos como este?

Un procedimiento alternativo, compatible con el etiquetado y que puede ser tratado conjuntamente, consiste en utilizar la *información visual intrínseca* de las imágenes (color y textura). Es en esta línea en la que desarrollaremos el experimento planteado en esta tesis.

El experimento básicamente consiste en aplicar un programa de ordenador que clasifica instancias –imágenes en este caso- bajo conceptos léxicos, es decir, conceptos equivalentes a una palabra, o a un conjunto reducido de palabras. Posteriormente se compara el rendimiento obtenido por el ordenador con el de los sujetos humanos que realizan la misma tarea.

¿Cómo funciona el programa? Primero, a partir de unas imágenes de entrenamiento clasificadas previamente por un agente humano, en relación a un concepto determinado, se obtienen los parámetros de una función probabilística (al conjunto de parámetros lo llamaremos 'vector beta'). Dicha función se aplica posteriormente sobre nuevas imágenes y nos da la probabilidad de que estas se incluyan bajo el concepto.

A modo de adelanto cabe reseñar que las nociones principales implicadas en nuestro experimento, explicadas con detalle en secciones posteriores, son:

- ***el vector beta***, que contiene los parámetros de la función -sobre un determinado concepto- que al aplicarla a los descriptores de una imagen nos indica la probabilidad de que dicha imagen sea instancia de dicho concepto;
- los conjuntos de ***imágenes positivas y negativas*** que sirven para aprender el concepto, o sea, ***la base de entrenamiento***;
- ***el centroide***, es decir, la imagen no real definida por la distancia mínima respecto a todas las positivas (centroide positivo), y lo mismo con las negativas (centroide negativo).

Los resultados obtenidos mediante el programa –o sea, los resultados del experimento– serán comparados con los que obtiene un sujeto humano. A partir de ahí se discutirá su pertinencia e implicaciones para el debate sobre la naturaleza y

el aprendizaje de los conceptos, planteando algunas cuestiones con gran tradición filosófica a sus espaldas, en particular, si un agente -humano o no– que es capaz de clasificar adecuadamente una serie de instancias bajo un concepto determinado, ha aprendido el concepto o no; qué consecuencias tiene esto, si es que tiene alguna, respecto a la estructura y el estatus ontológico de los conceptos;... Nuestro objetivo no es, desde luego, resolver estas cuestiones. Nos conformaremos con reenfocarlas y reflexionar sobre ellas sin ánimo de exhaustividad pero, eso sí, desde el ángulo que ofrece nuestro experimento. En este sentido esta tesis debe verse como una primera aproximación cuyo objetivo es prefigurar diferentes líneas discursivas, en relación al debate contemporáneo sobre los conceptos, que exploten los resultados del experimento y en las que profundizar en todo caso en trabajos posteriores.

Una computadora puede ser llamada ***inteligente*** *si logra engañar a una persona haciéndole creer que es un humano.*

Alan Turing

CAPÍTULO

2

El experimento

2.1 *Reconocimiento y clasificación de imágenes por ordenador*

El reto de nuestro experimento será implementar un programa informático (un algoritmo) en una máquina para que esta sea capaz de clasificar instancias (imágenes) bajo un concepto. Nuestra propuesta se articula como sigue: "A partir de unas imágenes catalogadas o etiquetadas con un determinado concepto –base de entrenamiento-, ser capaz de aprehender el concepto a partir de sus descriptores -calculados según los métodos expuestos en el siguiente apartado. Y, para que este aprendizaje sea válido, *evaluar su eficiencia en la predicción*, es decir, dada una imagen, identificar si es una instancia o no de ese concepto".

Un primer problema será cómo traducir la información visual de una imagen para que el ordenador, que no ve, sea capaz de discernir diferencias entre cada una de las imágenes en tanto que imágenes. El segundo problema es que, para que podamos decir que la máquina aprende, una condición necesaria es que pueda pronunciarse sobre imágenes nuevas. De algún modo, la máquina debe encontrar algún elemento o patrón común entre todos los casos positivos, y ser capaz de *extrapolarlo con éxito* a nuevos ejemplos, de manera que la presencia de dicho elemento se

convierta en criterio para estimar si la nueva imagen "cae" o no bajo el concepto. Dicho brevemente, la máquina debe generar una "hipótesis" acerca del concepto en cuestión basándose en las imágenes que se le suministran. Dedicaremos este capítulo a discutir ambos problemas.

2.1.1 *El tratamiento de la información visual contenida en las imágenes*

¿Cómo aprenden los conceptos los humanos? ¿Cómo aprendemos el concepto 'gato', por ejemplo? En el caso del aprendizaje humano -acotaré todos los sentidos de percepción en uno solo y sesgado, la vista bidimensional y estática- alguien nos tiene que mostrar una primera imagen de gato y decirnos que es un gato. Pero esto es solo una instancia de gato, es decir, no solamente esa percepción, esa imagen, es gato, sino que puede haber muchas más. Después vemos más imágenes de gatos, en las que alguien nos proporciona también la etiqueta 'gato'. Y con el tiempo no solamente se nos proporcionan instancias de 'gato', sino también de qué no es 'gato', para evitar confusiones y adquirir la habilidad de discriminar en casos más controvertidos.

Determinar por qué clasificamos el mundo del modo en que lo hacemos, por qué algunas de las cosas que existen son entidades enumerables y reidentificables (los perros) y otras no (piénsese en una sustancia como el agua, por ejemplo), a pesar de que ambas sean entidades perceptibles, es competencia de la neurociencia, la psicología cognitiva y también la filosofía (P. F. Strawson y W. v. Quine, serían ejemplos destacables en este último campo), que estudian cómo llegamos a elaborar una categorización de la realidad –una "ontología de conceptos", en la jerga de la inteligencia artificial. Pero ese no es asunto nuestro aquí. El problema para nosotros es cómo hacer que el ordenador "vea" las imágenes –las instancias– que sirven de base al concepto.

Supongamos, entonces, que ya contamos con un conjunto de imágenes clasificadas en instancias positivas y negativas por un agente humano. Esta será la "base de entrenamiento" para la máquina. Naturalmente, el ordenador no ve las imágenes, como puede verlas un sujeto humano. ¿Cómo transmitir entonces la información visual contenida en las imágenes a un ordenador?

El primer sesgo que debemos admitir es que la imagen real nunca será totalmente representada por la imagen digital. Este problema es común para todos los sistemas digitales, ya que la transformación de una señal analógica en digital, proceso conocido como digitalización, supone una primera cuantización del mundo real sobre un sistema numérico finito.

Dada una imagen digital, su representación informática suele estar definida como un mapa matricial de píxeles –que representan a todos los puntos discretos de la imagen-, donde en cada punto se captura el color mediante un esquema de los colores primarios RGB (de Red, Green, Blue). La cantidad de colores capaces de ser representados está en función de los bits dedicados a cada uno de los canales RGB, así hablamos de 255 tonalidades por canal cuando dedicamos 8 bits a su representación, es decir,

$$2^8 = 256 = 0, 1, 2, .., 255.$$

Hay representaciones de 16 bits, de 24 bits y de 32 bits. A esta última se le suele llamar en el argot informático "color verdadero", dada su pequeña granulometría.

Pero, siguiendo con esta estrategia, la cantidad de información que tenemos en cada imagen todavía es excesiva, y en muchos casos redundante, para extraer información semántica o "conceptual". De hecho, a partir de los estudios de física y medicina sobre la forma de procesar la información visual por parte del ser humano se sigue que la información procesada presenta una granulometría más gruesa (Rolls & Deco, 2001). Esto, junto con las dificultades inherentes al cómputo de excesivos datos, aconsejan una segunda cuantización.

En efecto, en esta segunda fase dejamos de mirar como una máquina e intentamos acoplarnos a los sistemas ópticos humanos. En primer lugar trasladamos el esquema de colorimetría de RGB a otro esquema de representación, el HSV, que es un espacio tridimensional con tres coordenadas con semánticas diferentes (matiz, saturación y brillo). Existen unas funciones que realizan esta transformación sin pérdida de contenido. Para computar los descriptores de color utilizamos, pues, el esquema de representación HSV de modo que a cada píxel de la imagen le corresponde un punto en este espacio. De hecho, la transformación de RGB a HSV (v. infra anexo 6.1.2, p. 141,) es biyectiva, con lo que se garantiza que no habrá sesgo

en la información transformada. Una vez realizada, cada uno de los píxeles de la imagen puede representarse por el valor de su punto en el espacio HSV.

En una fase ulterior se procede a la cuantización que dará lugar al nacimiento de los descriptores. Estos descriptores pueden ser agrupados según los parámetros físicos que intervienen en:

- Color
- Textura

Para obtener los descriptores, o características primarias, es necesario segmentar el espacio de representación mediante su partición en unos cubos determinados, y calcular la cantidad relativa de píxeles de cada cubo. Usualmente utilizamos el Matiz (H), cuyo arco son 360 grados, repartido proporcionalmente en cinco o más partes. La Saturación (S) en tres o más segmentos, y el Brillo (V) en dos o más. Por tanto tenemos como mínimo

$$5 * 3 * 2 = 30$$

descriptores en cada imagen.

Además utilizamos estos descriptores a nivel global, es decir, para toda la imagen, y también a nivel local, es decir, segmentamos la imagen en $\boldsymbol{n}$ trozos y calculamos sus descriptores como si cada segmento fuera una única imagen. Se suelen utilizar de 4 a 8 segmentos, con el fin de que no haya demasiados descriptores. Si tomamos 4 segmentos, estamos añadiendo

$$4 * 30 = 120$$

descriptores por imagen.

Para computar la Textura el proceso es más complejo. Utilizamos granulosidades e intentamos encontrar regularidades en la pixelación. Usualmente tomamos los grados 0 y 90, es decir, a nivel horizontal y a nivel vertical. Son funciones físico-matemáticas más complejas que aquí no comentaremos. De estas transformaciones obtenemos alrededor de 70 descriptores adicionales.

Finalizado el proceso, cada imagen queda representada por los valores del conjunto de sus descriptores, aproximadamente 220. [1]

[1] Hay otros descriptores de forma que funcionan con una semántica similar, pero cuyo proceso

Las cuestiones sobre las que deberíamos plantear la reflexión son: ¿realmente se está captando la realidad física de la imagen con este conjunto de descriptores? ¿Tiene un significado esta nueva realidad como lo tiene la imagen para un sujeto humano?

A la primera pregunta es fácil contestar que no: no se está captando la realidad física, si estamos pensando que la realidad se nos escapa en cuanto no aprehendemos toda la información. Pero, según este criterio, tampoco la estaría captando el aparato óptico humano. En primer lugar sufrimos el sesgo de la digitalización -una primera cuantización a la que nos referiremos como digitalización, que tiene en cuenta la posición dentro de la imagen-, y en segundo lugar, el sesgo de la cuantización -nos referiremos a esta segunda cuantización simplemente como cuantización, y que no tiene en cuenta la posición dentro de la imagen-. Mediante el procedimiento descrito anteriormente hemos pasado de la realidad visual a una realidad visual virtual, y de esta a una realidad numérica disminuida, con un importante sesgo. Quizá no debería preocuparnos el primer sesgo sufrido, ya que el ser humano también lo sufre; pero ¿qué pasa con el segundo, la cuantización? ¿hay una pérdida de información suficiente como para provocar una pérdida de representatividad? ¿también la sufrimos los humanos? En el segundo sesgo deberíamos plantearnos si una misma representación sirve para dos realidades visuales virtuales distintas, dado que sí es posible pensar en un escenario donde esto se pueda producir. Nótese que, por la forma de obtener los descriptores, un mismo valor de un descriptor de color (*bin*) no distingue: primero, las distintas tonalidades de su intervalo (H), ni de brillo ni de saturación; segundo, al ser el *bin* un recuento de píxeles de la imagen, no tiene en cuenta su posición en el espacio, así un punto rojo en la posición de la imagen (0,0) tiene el mismo valor para el descriptor que el mismo punto rojo en (230, 123); tercero, esto mismo aplicado a las particiones de la imagen para los descriptores de color local.

En este sentido, podemos afirmar que dos imágenes distintas pueden representarse de igual modo para el sistema, con lo cual serían indistinguibles. De todos modos, también el ser humano tiene dificultades en encontrar las diferencias entre

de extracción es extremadamente más complejo (Detectores de bordes de Harrison, SIFT, SURF, etc.) y que dan lugar a otros conjuntos de descriptores computacionalmente más costosos de obtener. Una comparativa de conjuntos de descriptores visuales se puede ver en (Deselaers et al., 2008)

dos imágenes similares. Como conclusión deberíamos aceptar este modelo como operativo, apto para abordar el problema, aun reconociendo sus limitaciones.

Podemos decir que el conjunto de descriptores tiene significado en tanto es una definición de la imagen atendiendo a sus características (físicas) o descriptores, aunque obviamente no significa lo mismo para un observador humano la imagen que el listado.

Llegados a este punto tenemos, pues, un mecanismo para representar la realidad visual mediante sus descriptores que, con las salvedades mencionadas, constituye un modelo válido de representación de esta realidad.

2.1.2 *El problema de la clasificación*

La estrategia apuntada al comienzo de este capítulo es exigente; de hecho, es similar a la puesta a prueba de una hipótesis científica. Así, primero, se ha de dar una explicación de por qué instancian un determinado concepto las imágenes así etiquetadas incluidas en el conjunto de entrenamiento. Entendemos 'explicación' aquí en un sentido laxo: la hipótesis explica en la medida en que *describe* un patrón, en cierto sentido compartido por esas imágenes precisamente (recuérdese la noción de centroide, v. supra cap. 1, p. 3, v. infra cap. 2.3.4, p. 47). Por eso la hipótesis no es explicativa en un sentido fuerte, es decir, no da cuenta de por qué existe dicho patrón común, simplemente lo explicita. Conviene advertir que el uso de explicación aquí es el acostumbrado en el contexto de la inteligencia artificial, sin embargo no es el usual en filosofía de la ciencia[2].

En segundo lugar, y para ser una hipótesis científica aceptable, se ha de contar con evidencia independiente a la utilizada para generar la hipótesis (v. por ej., Popper, 1994[1963] caps. 6, sec. 6 y cap. 10, sec. 5). En nuestro caso eso signifi-

[2]Como es bien sabido, el modelo de cobertura legal (covering-law model) propuesto por C.G. Hempel en los años cuarenta del siglo pasado, fue el primer análisis sistemático de la noción de explicación científica. Aunque actualmente hay acuerdo respecto a las limitaciones insalvables del modelo hempeliano, coexisten diversos enfoques (unificación, causal-mecanicista, etc.) respecto a qué cabe entender por explicación en la ciencia, sin que haya consenso respecto a ninguno de ellos. En cualquier caso, todos están bastante alejados del uso acostumbrado del término 'explicación' en inteligencia artificial. Este último es el que utilizaremos aquí para no comprometernos innecesariamente con alguna de las concepciones filosóficas de explicación científica en disputa. Una revisión clásica del debate post-hempel hasta finales de los años 80 se encuentra en (Salmon & Kitcher 1989). Una perspectiva más reciente puede encontrarse en (Psillos 2002).

ca que la hipótesis debe ser capaz de predecir un acontecimiento futuro, lo que en este contexto significa: dada una imagen sin etiquetar, decidir con éxito si tiene o no el concepto para el que la máquina ha sido entrenada. El vector beta, o sea los parámetros de la función que, aplicada sobre una instancia-imagen particular, nos da la probabilidad de que esta "caiga" bajo el concepto estudiado, será la hipótesis.

Centrándonos en un concepto, construir una hipótesis que explicara *in extenso* este modelo sería relativamente fácil: bastaría con almacenar *n* puntos dimensionales –siendo *n* las características (en este caso descriptores)– de cada imagen que contenga el concepto. De esta manera, cuando se presente una nueva imagen, comprobaremos si sus coordenadas coinciden con alguno de los puntos guardados. Si coinciden afirmaremos que la instancia cae bajo el concepto; si no, lo negaremos. Pero esta hipótesis solo respondería perfectamente ante las instancias ya conocidas; cualquier instancia que se presentara diferente de lo almacenado como positivo sería catalogada como negativa, aunque instanciara el concepto. Esta sería una forma de aprendizaje muy básica. Es dudoso incluso que quepa hablar aquí de "aprendizaje" ya que no es posible la extrapolación. Cuando se detectara un fallo de este tipo, un falso negativo (una imagen que realmente instancia el concepto cuando se ha predicho lo contrario), se podría modificar la base de entrenamiento incorporando el supuesto caso positivo al conjunto de instancias válidas o positivas, y reiniciar el proceso. Pero esta estrategia es tosca, poco eficaz, e incrementa su complejidad espacial y temporal según aumentan las imágenes. [3]

Otro método más útil sería intentar aprender qué tienen en común una serie de instancias catalogadas como positivas, y qué las hace diferentes de otra serie de instancias catalogadas como negativas. En este sentido ya no hace falta disponer de los puntos que caracterizan las instancias positivas, sino solo de una función que transforme un punto –de este hiperespacio R^n siendo *n* el número de descriptores– en su probabilidad de que contenga el concepto. Es decir, encontrar una función

$$f : R^n \rightarrow K,$$

[3] Entiéndase la complejidad espacial como la reserva de mayor espacio de memoria para almacenar los parámetros de las imágenes que deben ser guardadas por este modelo, y complejidad temporal como el incremento de tiempo en testear la cantidad de imágenes que conforman la hipótesis.

siendo K el conjunto de números reales del intervalo [0,1] que, dados los descriptores de una imagen, nos diga cuál es la probabilidad de instanciar este concepto.

Del conjunto de funciones que encajan en este perfil trabajaremos con la función de distribución logística (función p_i). Esta función, al aplicar el modelo logit (función g_i), dota a la función g_i de la propiedad de ser lineal con las diferencias logarítmicas de la función p_i para ajustar la predicción de si una imagen tiene o no un determinado concepto. Así, el modelo trata de maximizar los g_i y encontrar los parámetros beta, o vector beta, que los satisfacen (v. infra anexo 6.1.1, p. 139; Cap. 18, Russell & Norvig, 2010). El vector beta es la explicación del concepto, en el sentido de explicación propio a la Inteligencia Artificial ya indicado antes. Dicho vector, al ser aplicado a una imagen nueva, no incluida en la base de entrenamiento, mediante la función de probabilidad (p_i) -que no es exactamente una función de similitud- nos dirá la probabilidad de que esa imagen contenga este concepto. Estos parámetros conjuntamente con la función de distribución formarán la "hipótesis" del concepto en cuestión.

El primer paso, pues, será obtener para cada concepto la mejor explicación posible de acuerdo con la base de imágenes que se nos proporcione. Para cada concepto a entrenar/aprender, debemos tener dos subconjuntos de imágenes, uno con imágenes que satisfacen el concepto y otro con imágenes que no lo satisfacen. Entonces, para cada concepto, y utilizando los dos subconjuntos mencionados, ajustamos un modelo que maximiza la diferencia logarítmica de la probabilidad de que las imágenes que tienen el concepto sean positivas (lo más cercanas a 1) y las que no lo tienen sean negativas (lo más cercanas a 0). Este proceso se realiza computacionalmente usando la función logit. Su trabajo consiste en obtener el vector beta que mejor ajusta el modelo, dados los vectores X de descriptores de las características virtuales que se estén tratando de cada una de las imágenes de los dos subconjuntos de entrenamiento de este concepto (para mayor discusión matemática, v. infra anexo 6.1.1, p. 139).

Puede ocurrir, y de hecho ocurre a veces, que construida la hipótesis del concepto, al aplicarla sobre el mismo conjunto de entrenamiento, produzca falsos positivos –es decir, nos dice que una imagen instancia el concepto cuando ha sido etiquetada como que no– o falsos negativos –se dice de la imagen que no instancia el concepto a pesar de que la imagen pertenece al subconjunto de la base de entrenamiento

constituido por los casos positivos. Estaríamos ante un caso en el que el concepto no pueda ser explicado con los descriptores usados.

Una característica del modelo logit utilizado es que detecta combinaciones lineales de descriptores, es decir, si un descriptor depende como combinación lineal de otros, su parámetro dentro del vector beta será cero, indicando que su aportación es nula. Por ejemplo, supóngase que el subconjunto de instancias positivas para un determinado concepto contiene solamente dos imágenes caracterizadas por los vectores de descriptores (1, 1, 1) y la otra (2, 2, 2). En ese caso, nuestro modelo sería incapaz de encontrar más de un descriptor significativo para las imágenes positivas, luego la explicación estaría basada únicamente en un descriptor. En realidad, si solo le hemos proporcionado estas dos imágenes positivas, la hipótesis del concepto ha resuelto bien el problema, pasando a ofrecer la explicación más simple con una sola dimensión. La reducción dimensional en este sentido no debe percibirse como una pérdida de potencia explicativa, pues, sino más bien como la aplicación de la "navaja de Occam" a favor de la explicación más simple (argumento de la parsimonia).

Por otra parte, según aumentamos los ejemplos positivos, si se sobrepasa en mucho el número de descriptores, el modelo empieza a no converger, es decir, es incapaz de explicar los ejemplos suministrados por la base de entrenamiento. La prueba de que eso está ocurriendo la encontramos al darnos falsos positivos y/o falsos negativos sobre el conjunto de entrenamiento.

¿Cómo resolver, o minimizar, este problema? Aumentando el número de descriptores, parece el camino más directo. Pero, ¿y si físicamente no pudiéramos capturar más descriptores, es decir, si nuestra percepción (visual) del mundo físico estuviera agotada? En el caso humano aún podríamos recurrir a otros sentidos, como el oído, etc. Para el ordenador deberíamos "inventarnos" otros descriptores relacionados con la realidad física que no sean combinación lineal de los descriptores que ya tenemos. Estaríamos subiendo un escalón en la abstracción. [4]

[4]La solución a este problema queda abierta. Otros modelos utilizan funciones vectoriales que trasladan las n dimensiones a n+1 con el fin de encontrar el hiperplano n dimensional que divide en dos subespacios el hiperplano n+1 donde en un subespacio se encuentren las imágenes positivas y en el otro las negativas. Pero en este caso la dificultad radica en encontrar esta función y este hiperplano. Además, solo sería útil para explicar el conjunto de entrenamiento, faltaría ver su capacidad para predecir.

En estos momentos estudio varias vías para solucionar este problema. No podemos utilizar una

Lo anterior tiene que ver con los errores que el modelo comete con los propios ejemplos incluidos en la base de entrenamiento (lo que remite a su vez a la proporción entre el número de imágenes positivas de la base de entrenamiento y el número de descriptores). Pero los errores genuinos, por así decirlo, no son estos, sino aquellos que el modelo comete cuando se confronta con imágenes nuevas, con las no incluidas en la base de entrenamiento. Y es que el vector beta tiene una eficacia, tanto explicativa como predictiva, aproximada, desde luego. La comprobación de la eficacia predictiva se realiza calculando su éxito en la extrapolación de un criterio a un ámbito diferente de aquel para el que fue elaborado ad-hoc (la base de entrenamiento).

En todo caso, no cabe esperar que los objetos puramente formales que construye el ordenador encajen a la perfección con el concepto, o sea, que tales objetos repliquen exactamente la clasificación que realizaría un humano, o un grupo de humanos, en condiciones idóneas. La máquina comete errores, como los comete un humano. Otra cuestión es cuántos y en qué circunstancias. En secciones posteriores analizaré y compararé estas tasas de error. No abordaré, sin embargo, la capacidad de aprendizaje posterior del modelo, es decir, no me plantearé cómo modificar, si es que se puede, la hipótesis, o cómo generar otra hipótesis a partir de ella, según los errores genuinos –o sea, predicciones fallidas sobre imágenes nuevas- cometidos en la clasificación. Aunque sí se apuntarán posibles soluciones (v. infra 2.3.5, p. 48). Veamos a continuación el experimento con detalle.

2.2 *El experimento: un constructor de hipótesis sobre conceptos*

La rama de Inteligencia artificial que trata la clasificación está dentro del aprendizaje automático supervisado (Cap. 18, Russell & Norvig, 2010). Utiliza un método de inferencia inductiva para encontrar patrones compartidos en los ejemplos positivos y opuestos en los ejemplos negativos para generalizar una hipótesis. En cierto

combinación lineal de descriptores como otro descriptor ya que el modelo lo rechaza, así que exploro dos caminos: (i) conseguir nuevos descriptores agrupando descriptores básicos y obtener la probabilidad de que contengan el concepto de referencia; (ii) mantener la misma idea de agrupación, pero calculando la probabilidad sobre otros conceptos ontológicamente más simples (cuadrado, círculo, etc.).

sentido, esta propuesta es similar a la realizada por sir Francis Bacon en su tratado *Novum Organum* de 1620. Básicamente consistía en encontrar regularidades en las características de unas muestras de ensayo (tablas) que poseían el concepto que se quería explicar, y que además, estas no estaban en las muestras negativas. A partir de aquí se infería de manera inductiva que quien posee estas características es una instancia del concepto estudiado. El filósofo de la ciencia e historiador de matemáticas Donald Gillies, afirma que, aunque el método de Bacon se ha aplicado poco en la ciencia, al desarrollarse la rama de inteligencia artificial dedicada al aprendizaje automático, este método cobró una relevancia especial (Gillies, 1996). Esta técnica se puede utilizar en aquellos campos en los que dispongamos de ejemplos positivos y negativos de un determinado concepto a aprender. Las instancias deben estar descritas por un conjunto finito de características, además, el mismo número de características por cada instancia, diferenciándose las instancias entre ellas por el conjunto de valores que las definen.

El presente experimento, que cumple con los requisitos expuestos en el parrafo anterior, se enmarca dentro de los sistemas de recuperación de imágenes CBIR (Content-based Image Retrieval), realizado por quien presenta esta tesis, aunque sus raíces hay que buscarlas en la colaboración con el equipo de investigación de inteligencia artificial UNED-UV, y en las publicaciones resultantes (Benavent et al., 2010; Benavent et al., 2012; Benavent et al., 2013; Granados et al., 2011; Castellanos et al., 2011; Castellanos et al., 2012; de Ves et al. 2016). Su principal objetivo era reducir la brecha semántica que separa la descripción de una realidad en términos de descriptores de propiedades físicas de bajo nivel frente a la semántica que un humano percibe de esa misma realidad. Con otras palabras, ayudar en la solución a este problema, en este campo, consistía en reducir la dificultad en la comprensión de la información que el usuario percibe de las características de bajo nivel de los datos multimedia. Específicamente, en el caso de la recuperación de imágenes, la brecha semántica es la falta de correspondencia entre la información de las características visuales (por ejemplo, histogramas) y la interpretación de estos datos por un usuario en una situación determinada. Imágenes visualmente similares en términos de características de bajo nivel pueden ser muy diferentes en términos de significado.

2.2.1 *Diseño lógico*

El experimento pretende validar el modelo defendido en la presente tesis y consiste en generar un constructor de hipótesis para clasificar imágenes que, además, desempeñe esta tarea de un modo razonablemente eficaz en comparación con agentes humanos. En este sentido, la hipótesis construida por el modelo debe ser capaz de explicar los ejemplos facilitados para entrenar el sistema, y además debe ser capaz de predecir, dada una imagen, si esta se puede considerar o no una instancia de un concepto determinado. La ilustración 2.1, p. 17, muestra la lógica del experimento.

Para llevar a término este experimento necesitamos una base de datos de imágenes etiquetadas, que la describiré en el siguiente apartado. Después comentaré el diseño del experimento, concretando el contenido de cada una de sus fases y, para terminar el capítulo 2, interpretaré los resultados obtenidos.

2.2.2 *Fuentes de datos*

El presente experimento se ha realizado utilizando la base de datos de imágenes "concept-train" de la tarea *"Annotation"*, la edición de 2012 del concurso organizado por ImageClef (http://www.imageclef.org/2012/photo). Esta organización facilita un conjunto de imágenes a los equipos participantes para realizar las tareas propuestas para el concurso, cada año diferentes, con una fecha de entrega que generalmente es a finales de junio. Mayoritariamente son equipos de investigación en inteligencia artificial de distintas universidades a nivel mundial, aunque también participan compañías privadas como Google o Xerox.

La base de datos consta de 15.000 imágenes, un subconjunto de la base de datos MIRFLICKR-25K (ver anexo 6.5, p. 146), etiquetadas con un nombre abstracto y también con un campo para cada uno de los conceptos que nos dice si la imagen en concreto tiene o no este concepto. Para el año 2012 se facilitaron 94 conceptos agrupados en familias (p.e. para la familia *'timeofday'* se han propuesto tres conceptos *'timeofday_day'*, *'timeofday_night'* y *'timeofday_sunrisesunset'*). Para este experimento se han utilizado siete conceptos agrupados en dos familias (v. infra tabla 2.2, p. 24). Su elección ha sido arbitraria.

Respecto a la base de datos con la que vamos a trabajar conviene tener en cuenta

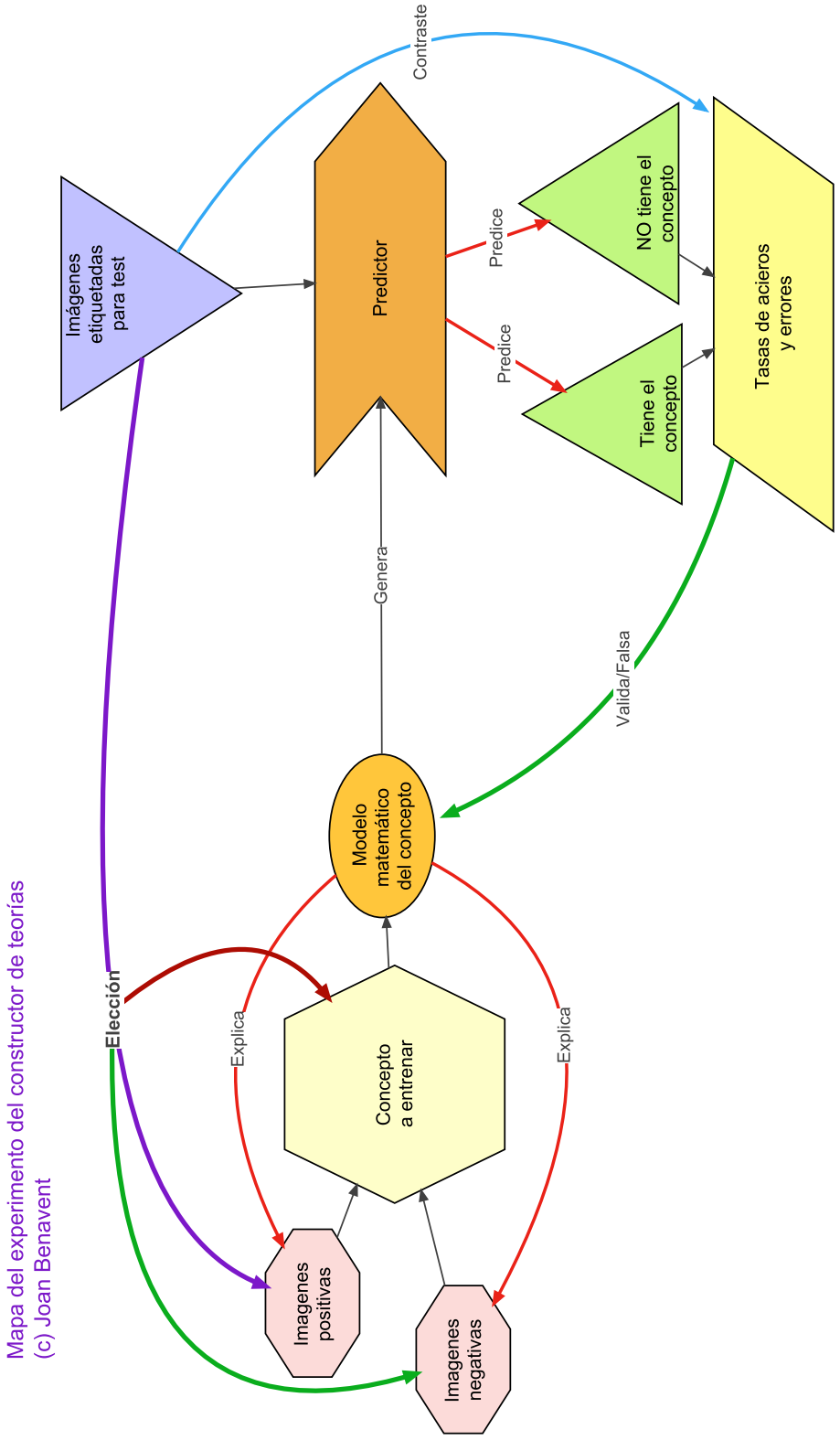

Figura 2.1: Lógica del experimento

lo siguiente:

- 1º) La forma de catalogar los conceptos por parte de la organización es un tanto peculiar. Consiste en ofrecer salarios a destajo a personas no especializadas, que son quienes deciden de modo individual si una imagen en particular ejemplifica o no un determinado concepto. Estos sujetos son los que aportan una clasificación inicial de las 15.000 imágenes, que es la que sirve de referencia para posteriormente seleccionar -en modo experto o en modo automático- una muestra de imágenes positivas y otra de negativas, así como para validar los aciertos y errores en la ejecución del experimento.

- 2º) Los conceptos catalogados sobre los que la máquina construye sus hipótesis son conceptos ”perceptuales visuales”, es decir, conceptos que pueden asociarse con imágenes sobre aquello a lo que refieren de forma relativamente directa. En este sentido, conceptos como ‘función’, ‘jurisprudencia’, ‘aleatoriedad’ o el propio concepto de ‘concepto’, por poner algunos ejemplos, no serían visuales, mientras que ‘perro’, ‘agua’, ‘sonrisa’,. . . , sí lo serían. No obstante, en un concurso de este tipo no se plantean conceptos visuales sencillos, donde pueda haber un consenso fácil entre humanos, sino que se exige una apreciación de matices que muchas veces pueden ser distintos en función del observador de la imagen. Compárese a este efecto el concepto ‘perro’con el concepto ‘timeofday_day’(que tal vez podríamos traducir como ‘horario diurno’).

Este planteamiento hace que, a menudo, la misma imagen sea clasificada de forma diferente, en relación a un determinado concepto, por más de un agente humano. Desde luego, esto no es un error por parte de los organizadores, pues lo que la prueba intenta reflejar es la complejidad existente en las bases de datos de imágenes que coexisten en las redes, aumentando así la dificultad del reto. Los detalles de la forma en que los agentes humanos han clasificado esta base de datos se pueden consultar en el anexo 6.5, p. 146, la semántica de los conceptos en el anexo 6.4, p. 145, y las tablas estadísticas de la clasificación realizada por humanos en 6.5, p. 148; 6.6, p. 148; 6.7, p. 148; 6.8, p. 148; 6.9, p. 149; 6.10, p. 149; 6.11, p. 149. De esta información, nos centraremos en que más de un 70 % de las imágenes clasificadas por agentes humanos bajo un concepto han sido obtenidas bajo el criterio de

un solo clasificador no necesariamente el mismo para todas las imágenes, y que por tanto, no ha estado sujeto a los estándares de calidad correspondientes a la plataforma informática. Por otra parte, para el concepto 0 "timeofday_day" los que sí han estado sometidos a este estandar, menos de un 30 %, en concreto el (29,8 %), de las clasificaciones, nos dan consenso entre los agentes solamente para el 59,24 %. Sobre estos últimos datos, o sea, sobre ese 29'8 % de imágenes que han estado sometidas a los estándares de calidad, es sobre lo que compararemos la eficacia de nuestro clasificador no-humano.

2.2.3 *Diseño físico del experimento*

El experimento consta de varias fases. En la primera se obtienen los descriptores de colorimetría y textura de las imágenes de la base de datos de entrenamiento facilitada, según el procedimiento descrito anteriormente en la sección (v. supra 2.1, p. 5). En la segunda fase se diseña la agrupación de descriptores para formar las características virtuales con las que trabajaremos cada regresión, es decir, cada modelo. En la tercera fase se eligen las imágenes con las que entrenaremos el sistema. En la cuarta se comprueba que el modelo explica las imágenes de entrenamiento suministradas y en la quinta ponemos a prueba el sistema suministrándole las imágenes de la base de datos para que realice predicciones sobre cada una de ellas. Por último, en la sexta fase analizamos las tasas de acierto y error en la explicación y predicción realizadas.

Fase 1: Obtención de los descriptores de colorimetría y textura.

Las imágenes suministradas en la base de datos están descritas en el espacio RGB, y necesitamos realizar una primera migración al espacio HSV, tal como se indica en los anexos matemáticos.

El siguiente paso es decidir cómo construiremos los descriptores de colorimetría. En este experimento se ha optado por realizar dos grupos de descriptores de color, el color general –ColorG- y el color local -ColorL-, el primero trata todos los píxeles de la imagen en el mismo sistema descriptor, mientras que el segundo segmenta la imagen a trozos y trata cada trozo como una misma imagen, yuxtaponiendo los descriptores obtenidos.

Para obtener los descriptores de ColorG hemos segmentado el espacio de re-

presentación mediante una partición de este en unos cubos determinados, y hemos calculado la cantidad relativa de píxeles de cada cubo, es decir, contamos los píxeles de un determinado cubo y lo dividimos por el total de píxeles de la imagen. La imagen 2.2, p. 21, muestra los cubos o bins en este espacio tridimensional.

Utilizamos el Matiz (H) cuyo arco son 360 grados repartido proporcionalmente en cinco partes. La Saturación (S) cuyo espacio es de cero a cien, en tres segmentos, y el Brillo (V), cuyo espacio es también de cero a cien, en dos. Por tanto tenemos

$$5 * 3 * 2 = 30$$

descriptores en cada imagen.

De esta forma, los píxeles de la imagen que se encuentren con un matiz entre 0 grados y 71 grados, saturación entre 0 y 32 y brillo entre 0 y 49, deben contabilizarse en el bin 1.

Para obtener los descriptores de ColorL, en primer lugar dividimos la imagen en 6 sub-imágenes, a cada sub-imagen le aplicamos el mismo mecanismo comentado para ColorG, a excepción de la forma de dividir los espacios HSV. En este caso, hemos dividido el matiz en cuatro partes, la saturación en cuatro y el brillo en dos. De esta forma, cada sub-imagen genera

$$4 * 4 * 2 = 32$$

descriptores, que componemos yuxtaponiendo los de cada sub-imagen, por lo que generamos

$$6 * 32 = 192$$

descriptores.

Para obtener los descriptores de granulometría utilizamos la textura en horizontal, la textura en vertical y la de línea base. Estas granulometrías se basan en evaluar los cambios de matiz, saturación y brillo en los ejes horizontal, vertical y diagonal. Se trata de encontrar patrones de repetición, y después pasarlos al espacio de frecuencias mediante la transformada de Fourier. Para el primero necesitamos 31 descriptores, para el segundo 31, y para el tercero 9. Se han eliminado los descriptores que no añadían información.

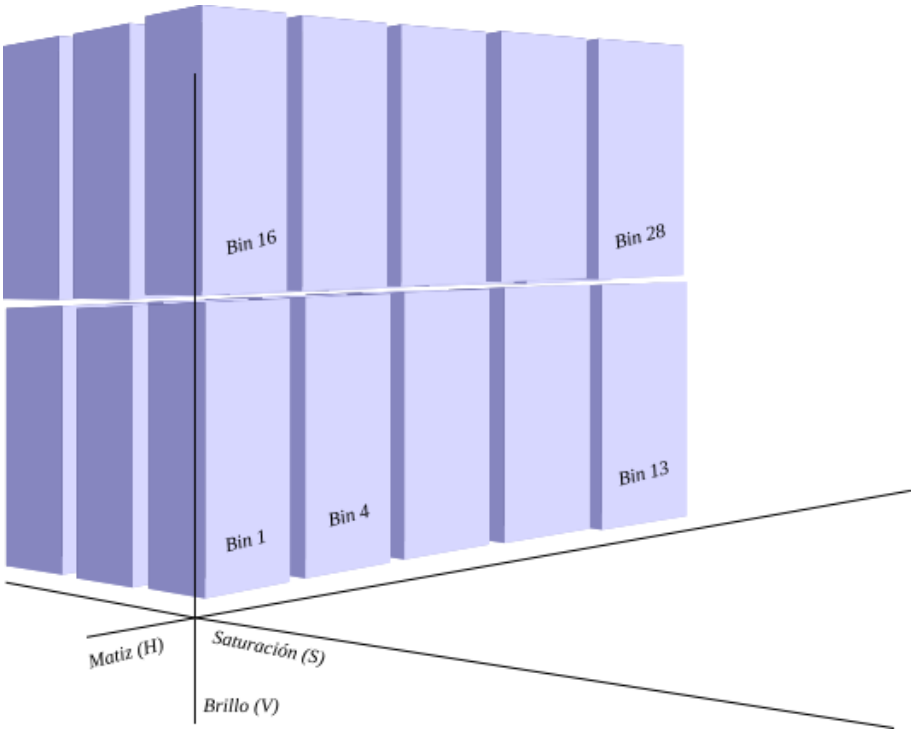

Figura 2.2: Modelo tridimensional de posicionamiento de los bins

Fase 2: Agrupación de descriptores para formar las características virtuales.

Los descriptores obtenidos en la fase anterior (293 d) los agrupamos para realizar las regresiones. De forma natural, ColorG tiene 30 descriptores, ColorL tiene 192, TextGral0 (granulometría horizontal) tiene 31, TexGral90 (granulometría vertical) tiene 31 y TextSsdl (granulometría diagonal) tiene 9. Según pruebas empíricas realizadas, el mejor comportamiento de las regresiones se ha observado cuando el número de descriptores por regresión es del orden del número de imágenes positivas suministrado, con un bagaje de hasta la mitad de descriptores (v. las publicaciones del autor y su grupo de investigación mencionadas en la bibliografía final). Intencionalmente, y tras algunos ajustes, se ha comprobado que para 20 imágenes positivas, 16 descriptores son suficientes.

De este modo, y teniendo en cuenta no mezclar en un mismo grupo descriptores heterogéneos, se han diseñado los siguientes grupos para las regresiones. A cada grupo le corresponderá una característica virtual. La tabla 2.1, p. 22, muestra cómo se agrupan los descriptores para formar las características virtuales.

Color G: 2 grupos
Color L: 12 grupos
TextGral0: 2 grupos
TextGral90: 2 grupos
TextSsdl: 1 grupo

Familia	Descriptores	Característica virtual
Color G (30 descriptores)	1..16	1
	17..30	2
Color L (192 descriptores)	31..46	3
	47..62	4
	63..78	5
	79..94	6
	95..110	7
	111..126	8
	127..142	9
	143..158	10
	159..174	11
	175..190	12
	191..206	13
	207..222	14
TextGral0 (31 descriptores)	223..238	15
	239..253	16
TextGral90 (31 descriptores)	254..269	17
	270..284	18
TextSsdl (9 descriptores)	285..293	19

Cuadro 2.1: Agrupación de descriptores en características virtuales.

De esta forma obtenemos 19 características virtuales, y a cada una de ellas le corresponderá un modelo, por lo cual hay que realizar 19 regresiones logísticas.

Fase 3: Elección de las imágenes que representarán a los conceptos. Imágenes de entrenamiento.

Sobre la base de datos etiquetada, tomaremos el siguiente criterio: *'una imagen ejemplifica un determinado concepto cuando algún humano así lo ha etiquetado y no lo tiene cuando ningún humano lo ha etiquetado'*. Ahora debemos seleccionar, de esta base de datos y con el criterio anterior, aquellas imágenes que representarán a cada concepto en concreto. Para tal fin se eligen veinte imágenes representativas del concepto (imágenes positivas) y sesenta que no representan al concepto (imágenes negativas).

El hecho de escoger el triple de imágenes negativas que de positivas tiene una importancia determinante para los conceptos tratados, ya que se necesitan ejemplos negativos para que el sistema reconozca qué imágenes no instancian el concepto a entrenar, y además, para estos conceptos, hay bastantes más imágenes, en principio, que no instancian el concepto que las que sí lo instancian, por tanto hay que suministrar más ejemplos negativos que positivos. Esta conjetura ha sido extraída a partir de la experiencia y los resultados obtenidos en varios años de investigación sobre el tema. (Benavent et al., 2010; Benavent et al., 2012; Benavent et al., 2013; Granados et al., 2011; Castellanos et al., 2011; Castellanos et al., 2012).

Hay diferentes formas de elegir las imágenes representativas de cada concepto, desde la elección manual por parte de un experto humano ("modo experto") que elige las imágenes que mejor representan el concepto como positivas, y las que peor, como negativas, hasta las más sofisticadas utilizando diferentes algoritmos matemáticos con medidas de similitud ("modo automático"). En nuestro caso el modo automático opera a partir de los centroides (v. infra 2.3.4, p. 47). Podemos definir el centroide en un hiperespacio de 'n' dimensiones, como el punto que minimiza la suma de las distancias euclídeas entre él y cada uno de los puntos facilitados para calcularlo. A partir del centroide de todas las imágenes positivas, se clasifican como instancias positivas las imágenes cuya distancia euclídea referida a las características virtuales quedan más cerca del centroide. El mismo proceso se aplica a las imágenes negativas, generando por su parte un 'centroide negativo'.

Para este experimento en concreto, trabajaremos con tres procedimientos o estrategias de selección de imágenes (*baselines*) diferentes, justamente para investigar los posibles efectos del modo de elección de las imágenes de entrenamiento sobre los resultados explicativos y predictivos.

Baseline1:

- Tanto las imágenes positivas como las negativas son seleccionadas **automáticamente**. Las primeras son las veinte imágenes positivas más cercanas al centroide formado por los descriptores de todas las imágenes etiquetadas como positivas en la base de datos (para un concepto determinado). Las negativas son las imágenes más cercanas al centroide formado por los descriptores de todas las imágenes etiquetadas como negativas y que **además no perte-**

nezcan a la misma familia[5] a la que pertenece el concepto dentro de la base de datos.

Baseline2:

- Las imágenes positivas son seleccionadas **manualmente** por un experto humano entre las imágenes etiquetadas como positivas para cada concepto en la base de datos, mientras que las negativas lo son **automáticamente**, tal como se ha especificado a propósito de Baseline1.

Baseline3:

- Las imágenes positivas son seleccionadas **manualmente** por un experto humano como se indica en el párrafo anterior, pero las negativas, elegidas **automáticamente**, serán las más cercanas al centroide formado por los descriptores de todas las imágenes etiquetadas como negativas que **además pertenezcan** a la misma familia a la que pertenece el concepto dentro de la base de datos. Con Baseline3 frente a Baseline2 se pretende comprobar el impacto de la familia en la elección de imagenes negativas, si es que lo tiene.

Por otra parte los conceptos a entrenar se muestran en la tabla 2.2, p. 24.

Concepto familia	Concepto número	Concepto nombre
Time of day	0	timeofday_day
	1	timeofday_night
	2	timeofday_sunrisesunset
View	70	view_portrait
	71	view_closeupmacro
	72	view_indoor
	73	view_outdoor

Cuadro 2.2: Conceptos del experimento.

[5]Para ver cómo se estructura la base de datos v. supra 2.2.2, p. 16 y p. 42; v. infra tabla 2.2, p. 24

Fase 4: Comprobación de que el modelo explica satisfactoriamente las imágenes de entrenamiento.

En esta fase realizaremos los ajustes necesarios para la convergencia del modelo. Se sabe que la regresión logística intentará aproximar las puntuaciones de las imágenes positivas a uno y las negativas al cero.

En este sentido podemos considerarla como una probabilidad, pero no hemos de confundirla con el concepto más clásico de probabilidad entre cero y uno, donde la equiprobabilidad es el 0,5. El modelo pretende separar las puntuaciones de las imágenes positivas hacia el 1, y las negativas hacia el cero. En este sentido, una imagen con una puntuación de 0,7 es más probable que sea una instancia del concepto que se está clasificando que otra imagen con puntuación 0,6. Ahora bien, esto no implica que un valor de 0,5 significa que es igual de probable que la imagen instancie o no instancie el concepto. Por tanto, podemos entender las puntuaciones más como medidas de similitud que de probabilidad. De todos modos, dejando a un lado la plausibilidad psicológica de la equiprobabilidad en un contexto como el que nos ocupa, nuestro modelo ofrece diversas alternativas para incorporar el caso de la imagen equiprobable (v. infra, p. 32).

Tal como se ha definido el experimento, se realizan para cada concepto diecinueve regresiones correspondientes a cada una de las características virtuales que hemos descrito. Cada característica obtendrá un modelo – una explicación– que intenta satisfacer o explicar la muestra suministrada, ofreciendo un resultado numérico cuando se le aplican los descriptores de las muestras al modelo. Como ejemplo veamos para el concepto 0 'timeofday_day', y la primera característica virtual que tiene dieciseis descriptores, aplicando los criterios de Baseline1 qué modelo beta de coeficientes nos proporciona:

β_{0-16} =(0.8031207, -19.46222, 5.742456, -2000.741, -2.407278, -11.63152, -571534.5, 28.41483, 453.7484, -4402753 , -41.49006, 24.67500, -1369.975, 31.8898, -5.615591, 2970.886, -24.95301)

Estos 17 valores, β_0 más los 16 parámetros restantes, al combinarlos con los valores de la característica virtual 1 de una imagen con la fórmula estadística de distribución usada por la regresión logística para hacer converger el modelo, nos proporcionará un valor numérico entre cero y uno.

Siguiendo el ejemplo, veamos qué puntuación correspondería a la imagen nu-

merada con -12360 y de nombre d589ec685fb2fa47f75645801aac2bd5.jpg, mostrada en la ilustración 2.3, p. 26, para esa característica virtual.

Figura 2.3: Imagen d589ec685fb2fa47f75645801aac2bd5.jpg

El valor de los descriptores de esta característica virtual es de:

$X_{1\text{-}16}$=(0.183877, 0.026816, 0, 0.050875, 0.012816, 0, 0.014069, 0.003728, 0 0.021451, 0.010629, 0, 0.234139, 0.027408, 0, 0.044485, 0.001488).

Al aplicar la función de *distribución logística*:

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \vec{\beta_i} * \vec{X_i})}}$$

Obtiene una puntuación de 0.9916178 para esta característica virtual. Se puede comprobar que al ser una imagen de las elegidas como positivas para la construcción del modelo, su puntuación es cercana al 1. Podemos afirmar, pues, que instancia el concepto 'timeofday_day'.

Veamos también para este mismo concepto, esta misma *baseline* o estrategia y esta misma característica virtual, qué puntuación obtiene una imagen de muestra negativa. Tomamos la imagen de nombre bc2b2bfc392a94d376cf253a7c3f38eb.jpg y numerada con -10743, mostrada en la ilustración 2.4, p. 27, para esa característica virtual.

Figura 2.4: Imagen: bc2b2bfc392a94d376cf253a7c3f38eb.jpg

El valor de los descriptores de esta característica virtual es de:

$X_{1\text{-}16}$=(0.094246, 0.001093, 0.000841, 0.020859, 0.000054, 0, 0.022324, 0.000342, 0, 0.055982, 0.000252, 0, 0.170517, 0.01642, 0, 0.181994)

Al aplicar la función de distribución logística se obtiene ahora una puntuación de 0.02985707 para esta característica virtual. Se puede comprobar que al ser una imagen de las elegidas como negativa para la construcción del modelo, su puntuación es cercana al 0. Podemos afirmar entonces que no instancia el concepto 'timeofday_day'.

En este sentido, las 19 regresiones ofrecerán 19 puntuaciones para cada imagen. Los 19 subvectores que integran este vector beta en esta baseline para el concepto 0 se muestran en el anexo 6.6, p. 150. Una forma sencilla de resolver la puntuación final, y de hecho es de las aproximaciones que mejor funciona en la práctica empírica, consiste en obtener la media aritmética de las 19 características y asignar esta puntuación a la imagen. Esta conjetura se ha probado en los experimentos y publicaciones del autor mencionadas en la bibliografía. Aplicando esta aproximación, los conjuntos de entrenamiento obtendrían las siguientes puntuaciones. Las positivas se pueden ver en la tabla 2.3, p. 28 y las negativas en las tablas 2.4, p. 29 y 2.5, p. 30.

identificador	Nombre	Puntuación
-12360	d589ec685fb2fa47f75645801aac2bd5.jpg	0,6475768
-11603	c9ebe36dd7a4a149e9e76548d08957.jpg	0,3359989
-11402	c6a9b7537694ca97f3c93c95411694.jpg	0,3328581
-10195	b338c6b784bcc73f85b24772c74519.jpg	0,4406566
-8949	9dfba01aaa80fc4a211a481c3b6682c9.jpg	0,3291097
-7876	8ce1dd225b87edf7aba80ea771a38fe.jpg	0,5205332
-7825	8c18e5beafa4c8594534fd389f46bef6.jpg	0,6570414
-7689	89b63d3d4d83b4f0f372dd22e64bb78f.jpg	0,6296838
-6968	7df284b1bc5f2cb3c7e421fe83a15c.jpg	0,5201109
-6614	7810b7cd9347f55874dabd382fb36a9f.jpg	0,5065418
-6164	70bb2669e5af83d0dd1ced5b2ba849.jpg	0,307076
-6119	701c21f9da3074ea553bce4199f488be.jpg	0,5166772
-6008	6e2515f0f7a2415044273018f1dd9.jpg	0,4132488
-6003	6dff4cdb0c6484534956e36e5e4025.jpg	0,3018918
-5844	6bbada97124ad36b5bc8dffba435bc60.jpg	0,4722252
-4028	4e9df87b37c93097a358b4f9b0172679.jpg	0,513251
-2711	39c9178a539eab341d42f96a34f3a5d.jpg	0,4951921
-1645	2954d9b363fc5967c91d2d273d9d66a.jpg	0,5304491
-1226	233dbcbed310a519fe2785357ea566b.jpg	0,5038265
-994 1	f3acd256f64ea399e7ccf814f7ec2.jpg	0,4071239

Cuadro 2.3: Imágenes positivas para concepto 0, con su puntuación para característica virtual 1.

identificador	Nombre	Puntuación
-14936	fed1bf661228dbc6aa51944ede8dc8.jpg	0,2871226
-14935	fed0eb3ae06858fa6695ea7ca247919.jpg	0,1483333
-14177	f3535971ed63605e499c1db22554025.jpg	0,3005906
-14051	f16b6f2f965cd6d39fc28643fc5b32.jpg	0,1597282
-13274	e58d6ed31241661e20535288cc44f2b.jpg	0,1804307
-12878	de39f63456cf59c2e7c6a132fcdbd91c.jpg	0,1268874
-12503	d7e02e8f32ece9a97f53734297d3a54b.jpg	0,2180819
-12402	d6365e879b2e983ff0a2fef0456745a1.jpg	0,1307792
-12373	d5c088271435d45f2a24b8908bfa488b.jpg	0,2355986
-12245	d38789413e177c9266355d81f51db7dc.jpg	0,163341
-11629	ca7aacc050f1a11e302383cf2ac23b.jpg	0,2356073
-11464	c7a7c3f279514eae8e9d895981ab89d.jpg	0,2178273
-11446	c76eed6fe3ef6b2c3d64aa9c31acb9a.jpg	0,133142
-11178	c33e746c8709a647e271be52d7bf79.jpg	0,1071475
-11091	c1c1c3866349d57be548775f9a63fdf8.jpg	0,2069889
-11037	c0fda1774e795bf36b91be95b1a44b13.jpg	0,1799342
-10898	beab9eaf84c5707827976af528f48d70.jpg	0,3357281
-10743	bc2b2bfc392a94d376cf253a7c3f38eb.jpg	0,1010234
-10687	baef25b2853660f42bb85cd2e95bc4f3.jpg	0,1732498
-10679	bad1133e5b5acad389e83c38daf77eed.jpg	0,1620971
-10472	b76ecb88fe1744294ba973c8e9c1bf53.jpg	0,2319089
-9986	afb2c482d3f87829521b249699e1f056.jpg	0,1103055
-9778	ac34605c42ff39443169f59642f24e.jpg	0,1350911
-9745	abb0c8691616a5b26958cd0903c239.jpg	0,1878244
-9546	a83831d9de9e1bbff27c87f842384992.jpg	0,1713133
-9502	a77c5a50f567c2fd8c0313b8eed4e51.jpg	0,1738076
-9045	9fbbad88c08411894dd9e2be86a29e8e.jpg	0,0931182
-8710	9a20f4eb125aff78339ff949a99027da.jpg	0,1322135
-8253	933a186f7ab0c6e2c0de407ff3762bea.jpg	0,1538281
-8152	91a4a866cea4d52da6efcc6475417b97.jpg	0,1697904

Cuadro 2.4: Imágenes negativas para concepto 0, con su puntuación para característica virtual 1.

identificador	Nombre	Puntuación
-8007	8f4dd3ae46b75fd7ddcd3975b9d7ae6.jpg	0,2370579
-7776	8b4727b59eedcdb0794bd912bcdff31.jpg	0,1394649
-7722	8a432b79c31b89f143653a63947dd6c.jpg	0,2117049
-7427	8558fa20c7bbef2d7b95d9991b2d71b.jpg	0,101312
-7346	83f8fa6d71ac32f6760ad1b46b5c6bf.jpg	0,1281482
-7303	835a1033f46c27a4ab2ed4f3bec5a927.jpg	0,1714942
-6898	7ccedac6218e035d88767ff021a625.jpg	0,1027398
-6884	7c83d44fb9b77a2b365a9cd707ae3.jpg	0,2276669
-6821	7b8857583bc1e6b78a91dd47f306fa3.jpg	0,2441433
-6645	7898b136706a7b239be2a222874e66a.jpg	0,1278544
-6494	761938e7f373a8ae9ba11cb31f53403a.jpg	0,1498026
-6379	742160056751d3f2c580824389cbf1.jpg	0,1283967
-5976	6daba628eb3630958e2f4c59a7eaa4e4.jpg	0,2679208
-5446	64d76b4191463fa0a53ca494f9edab5a.jpg	0,2451195
-5268	61dd30501c392e9a73ff5f5d5baf13.jpg	0,1038448
-5151	602533caf7573184764b63fc358123a5.jpg	0,1816823
-4737	59dae56ead4437af927ec62bbae87229.jpg	0,2322471
-4159	50aca41267ce74bfaec6f73cf124c96a.jpg	0,12584
-4122	50283b9a50e6a7da625c249e61a13f.jpg	0,204723
-4080	4f75d138d08c37e3bb390707352c4.jpg	0,1597826
-3836	4bb770b233abe99ede9de614ae504f.jpg	0,1418156
-3368	4475fe468911a6a9e17bd741c58a96c.jpg	0,1237017
-3332	43cbf1786aa9a2a790667f6125ff0a.jpg	0,2072072
-2745	3a5b614fd692c14e339ae21e40cabf5f.jpg	0,1613471
-2356	348a4bd7671836c4f51f7edd67e4f.jpg	0,1890529
-1886	2d8fbca2633827dbc9a539350d018.jpg	0,164568
-707	1ae19f3e69b4de2364bf726f4068037.jpg	0,136397
-563	18a40ba783e25bc1682605b7bdfab95.jpg	0,2250058
-545	184fb6b4a2dacf7f30c3708182c19acd.jpg	0,2199273
-259	137464be2ba0aa1e735962aa931ea4d.jpg	0,1961189

Cuadro 2.5: Imágenes negativas para concepto 0, con su puntuación para característica virtual 1.(bis)

Veamos cómo quedan las puntuaciones positivas y negativas, y si realmente existe una separación entre sus valores. La gráfica 2.5, p. 31, muestra esta separación de valores.

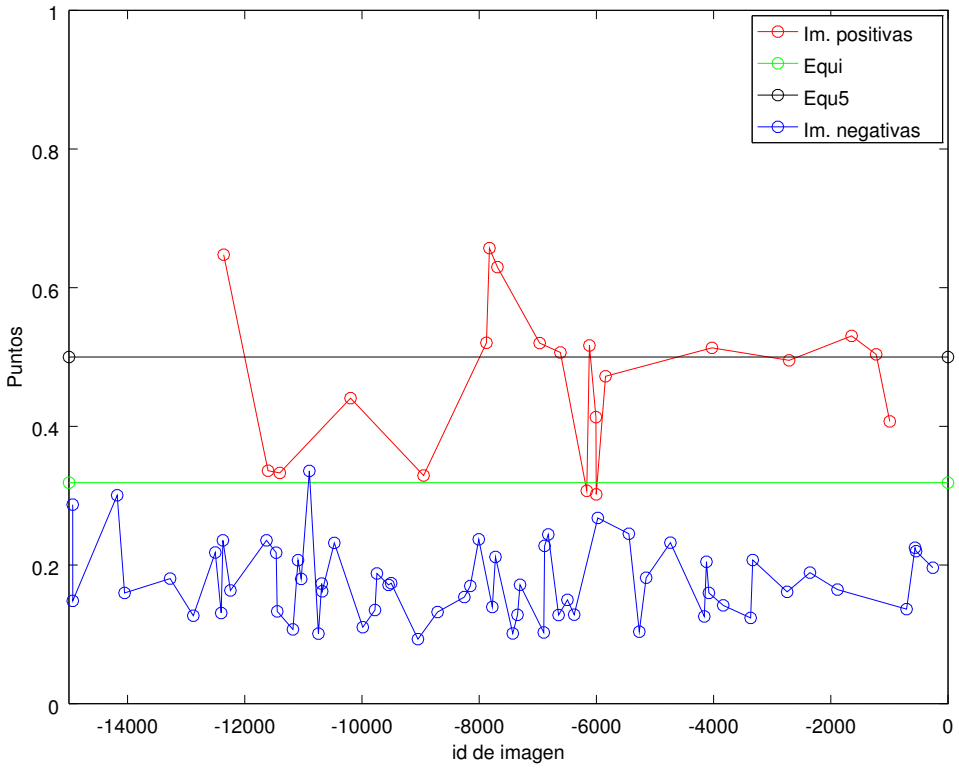

Figura 2.5: Puntuaciones de las imágenes de entrenamiento positivas y negativas.

Como ya se ha comentado antes, no podemos considerar esta puntuación entre 0 y 1 como una probabilidad, es decir, no podemos presuponer que un valor de 0,5 equivale a la equiprobabilidad de que esa imagen instancie determinado concepto. Para resolver este problema, es decir, para traducir la información cuantitativa -un número entre cero y uno- a un criterio dicotómico que permita discernir si se ejemplifica o no un determinado concepto, tenemos varias alternativas, las llamaremos "aproximaciones". Entre otras:

- 1ª) suponer que la equiprobabilidad se encuentra entre la mínima puntuación de las imágenes positivas y la máxima puntuación de las imágenes negativas

(optaremos por su punto medio);

- 2ª) definir la equiprobabilidad como el punto medio entre la media de las puntuaciones positivas y la media de las puntuaciones negativas;

- 3ª) igualar la equiprobabilidad a 0,5.

Las distintas formas de resolver este problema tendrán repercusión en la bondad de las explicaciones y en las predicciones del modelo. En este sentido, si tomamos la primera aproximación, la equiprobabilidad se encuentra en el valor 0,31880995; si tomamos la segunda, el valor es 0,3230178825; y si tomamos la tercera, la equiprobabilidad coincide con 0,5. En el presente experimento solo trataremos la primera y la tercera aproximación. Los valores para la equiprobabilidad según la aproximación primera se muestran en el anexo 6.3, p. 144.

La que mejor explica el modelo es la primera aproximación, ya que consigue explicar un 96,25 % con éxito, para el concepto 0 y baseline1. En concreto, de las imágenes positivas se explican 18 y se fallan 2, lo que supone un 90 % de acierto, mientras de las negativas se explican 59 y se falla 1, con una tasa de 98,4 % de acierto. La aproximación tercera solo explica el 87,5 % de las imágenes. Estos datos se muestran en la tabla 2.6, p. 32, para aproximación 1 y en la tabla 2.7, p. 33, para aproximación 3.

	Positivas	Negativas
Aciertos	18	59
Errores	2	1
Explica el	90 %	98,34 %
Total acierto	96,25 %	

Cuadro 2.6: Estadística de las imágenes de entrenamiento que explica el modelo para el concepto 0 para la aproximación primera en Baseline1.

	Positivas	Negativas
Aciertos	10	60
Errores	10	0
Explica el	50 %	100 %
Total acierto	87,5 %	

Cuadro 2.7: Estadística de las imágenes de entrenamiento que explica el modelo para el concepto 0 para la aproximación tercera en Baseline1.

Podemos concluir que el modelo –con la primera aproximación- explica satisfactoriamente las imágenes de muestra, para este concepto y para la Baseline1 (para más detalles, v. anexo 6.2, p. 142).

Fase 5: Predicciones realizadas por el modelo.

En esta fase, tomamos las 15.000 imágenes de la base de datos y para cada una de ellas realizamos predicciones sobre cada concepto. La distribución de los conceptos en la base de datos entre positivas y negativas por concepto se muestra en la tabla 2.8, p. 33.

Sin importar los posibles errores humanos al etiquetar las imágenes, y considerando que es ciertamente complicado definir un concepto con un número limitado de imágenes, las predicciones de nuestro modelo comparadas con las etiquetas de las imágenes en la Base de Datos se muestran en las tablas 2.9, p. 34; 2.10, p. 34 y 2.11, p. 35, para las tres Baselines en la aproximación 3, es decir, suponiendo la equiprobabilidad en 0,5.

Base de Datos			
Concepto	Total Im.	Im. Negativas	Im. Positivas
Concepto 0	15000	10103	4897
Concepto 1	15000	14315	685
Concepto 2	15000	14492	508
Concepto 70	15000	13467	1533
Concepto 71	15000	12660	2340
Concepto 72	15000	12939	2061
Concepto 73	15000	10144	4856

Cuadro 2.8: Estadísticas de las etiquetas de las imágenes de la BD para los conceptos tratados.

Baseline 1. Aproximación 3. (equprob. 0,5)				
Concepto	Acierto Positivas	Acierto Negativas	Acierto total	Acierto %
0	394	9570	9964	66,42666667
1	185	12867	13052	87,01333333
2	255	11156	11411	76,07333333
70	93	12541	12634	84,22666667
71	498	10676	11174	74,49333333
72	485	10886	11371	75,80666667
73	490	9490	9980	66,53333333
	2400	77186	79586	75,79619048

Cuadro 2.9: Estadísticas en predicciones. Aprox 3. Baseline 1.

Baseline 2. Aproximación 3. (equprob. 0,5)				
Concepto	Acierto Positivas	Acierto Negativas	Acierto total	Acierto %
0	1456	7301	8757	58,38
1	236	12011	12247	81,64666667
2	227	12021	12248	81,65333333
70	479	10051	10530	70,2
71	648	10179	10827	72,18
72	219	11841	12060	80,4
73	1305	8057	9362	62,41333333
	4570	71461	76031	72,41047619

Cuadro 2.10: Estadísticas en predicciones. Aprox 3. Baseline 2.

Baseline 3. Aproximación 3. (equprob. 0,5)				
Concepto	Acierto Positivas	Acierto Negativas	Acierto total	Acierto %
0	567	9165	9732	64,88
1	271	11015	11286	75,24
2	202	12094	12296	81,97333333
70	345	11146	11491	76,60666667
71	773	9592	10365	69,1
72	379	10934	11313	75,42
73	1073	8913	9986	66,57333333
	3610	72859	76469	72,82761905

Cuadro 2.11: Estadísticas en predicciones. Aprox 3. Baseline 3.

Suponiendo la equiprobabilidad en el punto medio entre la mínima positiva y la máxima negativa, (aproximación primera), cuya estrategia traslada más acierto en las explicaciones, nos encontramos con las siguientes predicciones en las tablas 2.12, p. 35, para baseline 1; 2.14, p. 36, para baseline 2 y 2.16, p. 37 para baseline 3, todas ellas predicciones globales. Para ver con detalle las predicciones distribuidas en positivas y negativas veamos las tablas 2.13, p. 36, para baseline 1; 2.15, p. 37, para baseline 2 y 2.17, p. 38 para baseline 3.

Baseline 1. Aproximación 1. Predicciones globales			
Concepto	Equiprobabilidad	Tot Aciertos	T. Ac. %
0	0.31880995	8648	57,7
1	0.32473215	7931	52,9
2	0.3432363	6767	45,1
70	0.3198159	7347	49
71	0.2905345	6237	41,6
72	0.3215002	6908	46,1
73	0.2971487	8513	56,8
		52351	49,9

Cuadro 2.12: Estadísticas en predicciones. Aprox 1. Baseline 1. Globales

Baseline 1. Aproximación 1. Predicciones parciales						
Concepto	Acierto Pos	Fallos Pos	Acierto %	Acierto Neg	Fallos Neg	Acierto %
0	2192	2705	44,8	6456	3647	64
1	489	196	71,4	7442	6873	52
2	408	100	80,3	6359	8133	43,9
70	873	660	57	6474	6993	48,1
71	1693	647	72,4	4544	8116	35,9
72	1392	669	67,5	5516	7423	42,6
73	2621	2235	54	5892	4252	58,1
	9668	7212	57,3	42683	45437	48,4

Cuadro 2.13: Estadísticas en predicciones. Aprox 1. Baseline 1. Parciales

Baseline 2. Aproximación 1. Predicciones globales			
Concepto	Equiprobabilidad	Tot Aciertos	T. Ac. %
0	0.14787856	5140	34,3
1	0.171522	2260	15,1
2	0.137789775	2040	13,6
70	0.170178195	2641	17,6
71	0.150273855	3437	22,9
72	0.12114043	2722	18,1
73	0.1578047	5447	36,3
		23687	22,6

Cuadro 2.14: Estadísticas en predicciones. Aprox 1. Baseline 2. Globales

Baseline 2. Aproximación 1. Predicciones parciales						
Concepto	Acierto Pos	Fallos Pos	Acierto %	Acierto Neg	Fallos Neg	Acierto %
0	4686	211	95,7	454	9649	4,5
1	664	21	96,9	1596	12719	11,2
2	491	17	97	1549	12943	10,7
70	1431	102	93,3	1210	12257	9
71	2157	183	92,2	1280	11380	10,1
72	1954	107	94,8	768	12171	5,9
73	4501	355	92,7	946	9198	9,3
	15884	996	94,1	7803	80317	8,9

Cuadro 2.15: Estadísticas en predicciones. Aprox 1. Baseline 2. Parciales

Baseline 3. Aproximación 1. Predicciones globales			
Concepto	Equiprobabilidad	Tot Aciertos	T. Ac. %
0	0.149393585	5967	39,8
1	0.22171294	3073	20,5
2	0.1506634	2871	19,1
70	0.10495042	2352	15,7
71	0.163734325	3144	21
72	0.197789615	3770	25,1
73	0.159277215	5757	38,4
		26934	25,7

Cuadro 2.16: Estadísticas en predicciones. Aprox 1. Baseline 3. Globales

Baseline 3. Aproximación 1. Predicciones parciales						
Concepto	Acierto Pos	Fallos Pos	Acierto %	Acierto Neg	Fallos Neg	Acierto %
0	4079	818	83,3	1888	8215	13,2
1	638	47	93,1	2435	11880	16,8
2	480	28	94,5	2391	12101	17,8
70	1445	88	94,3	907	12560	7,2
71	2196 1	44	93,8	948	11712	7,3
72	1723	338	83,6	2047	10892	20,2
73	4549	307	93,7	1208	8936	1,4
	15110	1770	89,5	11824	76296	13,4

Cuadro 2.17: Estadísticas en predicciones. Aprox 1. Baseline 3. Parciales

En un primer análisis parecen mejores resultados globales cuando se utiliza la aproximación tercera y la Baseline1, consiguiendo más de un 75 % de aciertos globales para todo el experimento. Por otro lado, la aproximación primera, que era la que mejores resultados ofrecía en la fase de explicación, se queda muy lejos de esas cifras, ofreciendo su mejor versión para Baseline1 con una tasa de aciertos que no llega al 50 %.

Fase 6: Análisis de las tasas de acierto y error.

Las estadísticas empíricas de aciertos y errores se pueden consultar en el anexo 6.2, p. 142, de la presente tesis. No obstante, los cuadros siguientes resumen lo más importante sobre la efectividad explicativa, tabla 2.18, p. 39, y predictiva, tabla 2.19, p. 39, de las distintas estrategias.

Si tomamos las mejores tasas de acierto global como factor discriminante de la bondad de nuestro modelo, o sea, contando tanto los aciertos en las imágenes positivas como en las negativas, encontramos que las mejores explicaciones para todos los conceptos las obtiene Baseline1 con la aproximación primera. Para las predicciones, los mejores resultados son para la aproximación tercera con Baseline1, con algunos conceptos (2, 73) que los obtienen con Baseline3.

¿Qué conclusiones generales podrían sacarse a partir de estos resultados?

- 1º) Los datos obtenidos confirman que ***el modelo explica mejor que predice***, en todas las baselines y para todas las aproximaciones. Esto es, en principio,

Mejores Explicaciones			
Concepto	Baseline	Aproximación	Tasa global de aciertos %
0	1	1	96,25
1	1	1	100
2	1	1	100
70	1	1	97,5
71	1	1	96,25
72	1	1	96,25
73	1	1	92,5

Cuadro 2.18: Mejores explicaciones de cada concepto

Mejores Predicciones			
Concepto	Baseline	Aproximación	Tasa global de aciertos %
0	1	3	66,43
1	1	3	87,01
2	3	3	81,97
70	1	3	84,23
71	1	3	74,71
72	1	3	75,80
73	3	3	66,57

Cuadro 2.19: Mejores predicciones para cada concepto

un resultado esperable, entre otros factores porque la hipótesis -modelo- se ha desarrollado a partir de la base evidencial suministrada al sistema, y en este sentido, debería ser capaz de explicar los ejemplos de entrenamiento al 100 %, cosa que no siempre hace. Cuando no lo consigue decimos que el modelo no converge, es decir, con los descriptores suministrados no ha sido capaz de encontrar un vector beta que separe las puntuaciones positivas de las negativas. En estos casos devuelve el vector beta que, para el conjunto de entrenamiento, más aproxima al uno las imágenes positivas y al cero las negativas.

- 2º) Tanto en la explicación como en la predicción, ***la forma de elegir las imágenes*** que sirven de ejemplo, es decir, aquellas a partir de las cuales se construirá el modelo, ***influye notablemente sobre los resultados***. Esto también parece natural que sea así, ya que el conocimiento se extrae del modelo mediante "aprendizaje supervisado".[6] En este sentido, Baseline1, la estrategia en la cual el agente humano no interviene en la selección de imágenes, parece funcionar mejor tanto en explicaciones como en predicciones. La forma de elegir los ejemplos positivos ha sido escoger las imágenes más cercanas al centroide formado por los descriptores de todas las imágenes etiquetadas como positivas en la base de datos para un concepto determinado, y por esta razón, Baseline1 tiene una información intrínseca de la distribución de estas imágenes en el hiperespacio generado por los descriptores virtuales. Las imágenes negativas se eligen de forma parecida, con la restricción adicional de no pertenecer a la familia a la que pertenece el concepto -Baseline1 y Baseline2-, o pertenecer a la familia del concepto -Baseline3-. Los resultados obtenidos mediante la selección automática de las imágenes positivas -Baseline1- superan a los obtenidos mediante la selección manual de un humano, recuérdese que eso ocurre en Baseline2 y Baseline3.

  Aunque podamos pensar que, en principio, estas últimas apelan a una información semántica del concepto más profunda, la incorporada en los veredictos de un sujeto humano que no se guía exclusivamente por las propiedades

[6] El aprendizaje supervisado es una técnica utilizada en inteligencia artificial para construir modelos a base de ejemplos catalogados suministrados a la entrada. Suele utilizarse en entornos de clasificación (Russell & Norvig , 2010).

físicas de la imagen, también es verdad que esta información depende del actor humano que realiza la selección, y que por eso puede haber un componente subjetivo; además, no tiene en cuenta el contexto particular en que se plantea el experimento, esto es, una base de datos, compuesta por imágenes. Por otro lado, el hecho de que para los conceptos 2 y 73 la mejor predicción sea la realizada por Baseline3, queda ensombrecida porque en ambos casos los resultados que obtiene Baseline1 son muy próximos -ver anexo 6.2, p. 142-, lo que no desvirtúa la afirmación general realizada sobre Baseline1.

- 3º) ***La aproximación primera es la que mejor explica, mientras que la tercera es la que mejor predice***[7]. Este hecho es justificable si tenemos en cuenta que la distribución de cada concepto en la base de datos global, compuesta por miles de imágenes, no es simétrica, es decir, que para cada concepto hay sustancialmente más imágenes negativas que positivas v. supra 2.8, p. 33. Por tanto, cuanto mayor sea el espacio entre cero y uno etiquetado para que una determinada imagen sea considerada negativa –y eso es lo que hace la aproximación tercera al igualar la equiprobabilidad a 0,5–, más aciertos "generales" tendremos[8]. Con otras palabras, ante la duda, con una predicción negativa ('la imagen no instancia el concepto') será más probable acertar que con una predicción positiva. Este debate, si se quiere mantener la equiprobabilidad en 0,5 conduce a plantearse la adopción de medidas ponderadas para contabilizar los errores, es decir, el coste de un fallo estará en función de la distribución estadística de la base de datos. ¿Quiere esto decir que necesitamos saber 'a priori' la distribución entre imágenes positivas y negativas de

[7] Recuérdese que la aproximación primera significa que el punto de corte/criterio de decisión es la media entre la puntuación más alta de las negativas y la más baja de las positivas, y que para la aproximación 3 el punto de corte = 0,5.

[8] La explicación de este argumento lo encontramos en el número de imágenes positivas y negativas que instancian un determinado concepto en la base de datos. Hay pocas positivas frente a las negativas v. supra 2.8, p. 33. Si suponemos - que no lo es - una distribución uniforme de las imágenes positivas y negativas entre cero y uno, se cometerán menos errores cuanto mayor espacio de acierto dejemos para el conjunto mayoritario. En este caso, las imágenes negativas son el conjunto mayoritario, de orden mayor de 10, por tanto, disponer de un espacio [0, 0.5] para catalogarlas como negativas presentará una tasa de errores menor que disponer de un espacio [0, 0.318] que presenta la aproximación 1 para baseline 1. Además, la mayor tasa de aciertos en positivas si su espacio fuese [0.318, 1] no compensaria el incremento de errores en negativas, por esta descompensación en la densidad de imágenes positivas frente a negativas en el intérvalo. Esto explicaría por qué predice mejor la aproximación 3.

la base de datos? Sí, eso parece. Entonces, ¿estamos seguros de estar aprendiendo algo distinto a meras estadísticas?

Además, la causa justificada del deslizamiento hacia abajo de la equiprobabilidad la encontramos en cómo se optimiza la regresión logística para obtener los parámetros del vector beta (v. infra anexo 6.1.1, p. 139).

Supongamos que escogemos tres veces más imágenes negativas que positivas para extraer los parámetros del vector beta de un determinado concepto -este es nuestro planteamiento en el experimento-. Sea i la cantidad de imágenes positivas y j la de negativas, con $j = 3i$. Cuando intentemos maximizar la función

$$\prod_{i=1}^{k} g_i * \prod_{j=1}^{J} g_j \text{ donde } i \in positivas \wedge g_i = \log\left(\frac{p_i}{1-p_i}\right) \wedge j \in negativas \wedge g_j = \log\left(\frac{1-p_j}{p_j}\right)$$

el coste de acercar a cero una instancia negativa es menor que el coste de acercar a uno una instancia positiva, porque la función a maximizar es un productorio de la función g_i de cada instancia positiva y g_j de cada instancia negativa. Téngase en cuenta que la función g_j tiene los argumentos del logaritmo invertidos respecto de la función g_i. Esta inversión hace que la optimización de una instancia positiva hacia uno sea valorada numéricamente igual que la optimización de una instancia negativa hacia cero para la función optimizadora.

- 4º) No se evidencian diferencias sustanciales en la tasa de aciertos entre Baseline2 y Baseline3, mientras que sí entre estas y Baseline1. Este hecho apunta a que en ***la elección de imágenes de entrenamiento negativas no importa la familia del concepto sino el concepto.***[9]

[9]Recuérdese que la diferencia en la elección de imágenes de entrenamiento entre baseline 2 y baseline 3 consiste en la forma de elegir las imágenes negativas: baseline 2 las busca entre las que ni tienen el concepto ni pertenecen a la misma familia de toda la base de datos, mientras baseline 3 entre las que no tienen el concepto pero pertenecen a la misma familia.

2.3 *Elementos relevantes*

En las secciones anteriores, se ha descrito con detalle el experimento de clasificación que dará soporte técnico a esta tesis.

Esta sección describe los elementos del experimento candidatos a ser piezas fundamentales de acople para poder discutir las diferentes teorías cognitivas sobre conceptos. Con otras palabras, los *'elementos del experimento relevantes para el encuadre conceptual'*.

Por una parte, trataremos los elementos derivados directamente de la percepción de la imagen, con pequeñas transformaciones vehiculares, como las imágenes en sí mismas, su transformación a los espacios RGB y HSV, la cuantización como un cambio hacia el espacio de los descriptores, las características virtuales, los centroides como representantes abstractos en el espacio de los descriptores y los conjuntos de imágenes elegidas para entrenar los mecanismos de aprendizaje del autómata.

Y por otra parte, veremos los resultados del proceso de aprendizaje del autómata, representados por dos elementos fundamentales, el vector beta y el límite de equiprobabilidad.

Finalmente discutiremos brevemente el símbolo lingüístico.

2.3.1 *Los descriptores de imagen*

En este apartado trataremos los descriptores de imagen como referencias a la imagen y la irreversibilidad de la transformación del espacio HSV a los descriptores de imagen.

Como ya hemos visto en el experimento, la imagen digitalizada se representa por una matriz numérica en el espacio RGB. Esta matriz es el vehículo para la representación de la imagen dentro del ordenador, su extensión suele ser (.jpg). Seguidamente se ha transformado al espacio HSV, como otra matriz numérica sin perdida de información, y además como una transformación biyectiva, es decir una transformación completamente reversible (v. anexo 6.1.2, p. 141). El formato de esta nueva matriz es directamente transformable en imagen por casi todos los programas informáticos de imagen.

Tanto la matriz RGB como la HSV son los vehículos de la imagen dentro del ordenador, sin perdida de información, es decir, su codificación interna permite representar en una pantalla la imagen original, hasta cierto grado de precisión, tal como se discutió en secciones anteriores.

A partir de la matriz HSV se procede a su cuantización para extraer los descriptores de colorimetría. Por otra parte, de la matriz RGB se extraen los descriptores de textura. Ambos, los descriptores de colorimetría y los de textura, conforman un vector numérico de 293 elementos que será el que represente a la imagen en los procesos de aprendizaje y clasificación.

Este vector de descriptores – de rango mucho menor que la imagen original en cualquiera de los espacios de representación RGB o HSV – que será el nuevo representante de la imagen, tiene unas características muy diferentes a las que tenían las matrices RGB o HSV. En efecto, se puede observar que sus dimensiones se han reducido notablemente. Hemos pasado de una matriz de tres columnas de anchura, tres columnas de profundidad (RGB o HSV) y, en función de su resolución hasta de más de mil filas, a un vector con menos de 300 elementos. Como ya se comentó en secciones anteriores, el proceso de cuantización es irreversible, es decir, debido a la pérdida de información en el proceso, no podemos obtener la matriz RGB o HSV a partir del vector de descriptores. Esta característica nos debe hacer pensar en un primer paso hacia una clasificación intrínseca al proceso de cuantización, es decir, el vector de descriptores representa a la imagen cuantizada, pero también a toda una serie de imágenes diferentes que tras su cuantización arrojan el mismo vector. Estos procesos ya se comentaron en la parte del experimento (v. supra 2.2.3, p. 19).

Se argumentó en secciones anteriores que esta pérdida de información es perfectamente asumible, y además es necesaria para iniciar el proceso de aprendizaje, ya que si consideramos toda la información el problema de clasificación se hace intratable computacionalmente, y además, este exceso de datos tampoco aporta información relevante al proceso.

En resumen, podemos ver el vector de descriptores como una referencia a la imagen que representa en el espacio RGB o HSV, y que además esta referencia admite más de una interpretación, es decir, admite ser representante de *más de una imagen en el espacio RGB o HSV*. Estos hechos provocan que el proceso de transformación inversa no sea único, con lo cual un mismo vector en el espacio de

los descriptores admita más de una interpretación en el espacio RGB o HSV.

Estamos ante una arquitectura a capas, donde el nivel físico, representado por las matrices RGB y HSV suponen la codificación más próxima a la realidad y más extensa en contenido. En la siguiente capa, que podríamos definir como el nivel de descriptores, la imagen es representada vehicularmente por el vector de descriptores. Todo indica que hemos subido un nivel de abstracción de la realidad, y trabajando en este nivel, sin perder la referencia a la imagen real, no podemos afirmar que el vector de descriptores sea ni una imagen real ni algo equivalente a ella, más bien es una representación de una imagen real en un formato simplificado (en un vehículo reducido), que serviría para representar otras imágenes reales.

2.3.2 *Las características virtuales*

Centrándonos en el nivel de descriptores, donde el vector de 293 elementos es el vehículo que representa a la imagen, todavía podemos distinguir elementos diferenciales dentro de su estructura. Por una parte tenemos la procedencia de los descriptores, los que proceden de características de colorimetría y los que proceden de características de textura. Por otra parte, y dentro de cada subdivisión, tenemos que dentro de los descriptores de colorimetría podemos distinguir si en los parámetros para su obtención se ha tenido en cuenta toda la imagen, o sólo una parte, de esta forma tenemos la familia ColorG y ColorL, y dentro de ColorL, todavía podemos distinguir cómo se ha hecho la partición de la imagen global. Dentro de la subdivisión de textura, también distinguimos la dirección que hemos seguido para encontrar patrones de textura, si hemos trabajado horizontalmente, es decir con grado 0, verticalmente, con grado 90, o diagonalmente, con grado 45.

Por otra parte, hay que ajustar tanto las imágenes de entrenamiento como la cantidad de descriptores que formarán parte de la regresión logística para que esta sea lo más eficiente. Tal como se discutió en el experimento, se optó por utilizar grupos de descriptores para cada regresión con un máximo de 16, y además exigimos que los descriptores de cada regresión pertenezcan a una misma familia, es decir que su extracción haya sido realizada de forma homogénea. Para más discusión ver el apartado dedicado a la agrupación de descriptores para formar características virtuales, (v. supra 2.2.3, p. 21).

Todas estas especificaciones de diseño nos llevan a considerar otro objeto abstracto a tener en cuenta, las características virtuales. En efecto, una característica virtual, en este sentido, no es más que el grupo de descriptores que forman parte de una misma regresión logística. De esta forma, siempre podemos ver el grado de similitud atendiendo individualmente a una única característica. Por otra parte, he decidido llamarla característica virtual porque no tiene ningún significado especial esta forma de elegir los descriptores, es decir sería muy difícil encontrar una semántica especifica -un concepto perceptivo ”natural”, que pueda asemejarse a aquellos que forman parte del esquema conceptual de los humanos- asociada a cada grupo de descriptores que nos pudiera ser útil en un futuro para la posible construcción de explicaciones semánticas mediante redes de inferencia.

Como ya se ha visto en la parte del diseño físico del experimento, hemos realizado 19 regresiones independientes en cada proceso de aprendizaje de un determinado concepto, por tanto, tenemos 19 características virtuales.

Si seguimos optando por una arquitectura a capas, el siguiente paso de abstracción nos conduciría a la capa de las características virtuales. En este sentido, el vector de descriptores podría definirse en términos de sus características virtuales. Este paso de abstracción no modifica el contenido del vehículo que referencia a la imagen (el vector de descriptores), ya que, aunque en este nivel de abstracción sólo tenemos 19 características, éstas son en realidad subvectores de descriptores que al yuxtaponerlos nos devuelve el mismo vector de descriptores como referencia a la imagen que representa.

Las posibilidades que pueden extraerse de tratar esta capa no han sido tenidas en cuenta en la presente tesis, y pueden ser objeto de investigación en un futuro.

2.3.3 *Las imágenes de entrenamiento.*

El experimento se basa, como hemos visto, en aprendizaje supervisado, es decir, se deben suministrar instancias clasificadas para aprender de ellas. Los conjuntos de imágenes positivas y negativas son estas instancias.

Básicamente se han experimentado dos formas de elegir estos conjuntos. Primero, que las imágenes sean elegidas por un experto humano, al que se supone que será capaz de aportar información semántica que no puede aportar la máqui-

na, y que, por consiguiente será más eficiente el aprendizaje. Y segundo, que las imágenes sean elegidas automáticamente como las más cercanas al centroide de las imágenes positivas, para positivas, y lo mismo para las negativas, aunque con distinción entre imágenes que no contuvieran ningún concepto de la familia (baseline 1 y baseline 2), y otro dentro de la familia (baseline 3). Como ya hemos visto, los conjuntos construidos automáticamente han demostrado ser más efectivos que los elegidos por un experto humano. Para más detalles consultar el apartado 2.2.3, p. 22.

Estos conjuntos de imágenes juegan un papel central en la construcción del modelo de clasificación, ya que de ellas se extrae el conocimiento para la clasificación. Cualquier variación en estos conjuntos supone una variación en el modelo de clasificación.

En este sentido, el conjunto de imágenes positivas podría entenderse como las imágenes que representan el concepto que se pretende aprender. Las imágenes negativas tienen un sentido de rechazo del concepto, "lo que no es", y sirve principalmente para afinar en la discriminación de lo que realmente instancia el concepto.

En resumen, el concepto que se pretende aprender puede ser representado por el conjunto de imágenes positivas y negativas, y en este sentido, será muy importante para la semántica del mismo la forma de elegir las imágenes. Concretamente dibujará escenarios distintos de aplicación del mismo concepto – se discutirá más a fondo en el capítulo 4, concretamente dentro del apartado sobre el neo-empirismo, apartado 4.1, p. 104.

2.3.4 *El centroide*

Tanto si las imágenes de entrenamiento han sido elegidas automáticamente o por un experto, se puede calcular un punto que representa el lugar central del conjunto de imágenes dentro del hiperespacio que describen los descriptores (293 dimensiones) considerándolo de geometría euclídea cuya métrica responde a:

$$ds^2 = \sum_{i=1}^{293} di^2.$$

definimos este punto, representante de una imagen abstracta, como el centroide – aquel punto que minimiza la suma de distancias euclídeas del conjunto de imágenes positivas o negativas-.

Este vector de descriptores, generalmente, no representa a ninguna de las imágenes de su conjunto, ni tenemos posibilidad de visualizar a qué imagen puede representar – por la irreversibilidad del proceso de cuantización discutido anteriormente -, pero es el mejor candidato a ser el representante de clase, es decir el representante abstracto del concepto en el hiperespacio de los descriptores.

Tomado de esta forma, se puede medir la distancia métrica entre dos conceptos sin más que medir la distancia entre sus dos representantes abstractos en el hiperespacio de los descriptores, es decir, sus propios centroides positivos. También es posible medir la distancia métrica entre los centroides negativos. Otra cuestión bien distinta es dotar de sentido a estas distancias en los términos que lo haría un humano como distancia psicologica entre conceptos. Estas cuestiones no serán tratadas en la presente tesis.

2.3.5 *El vector Beta y la equiprobabilidad*

Sin duda, el elemento más interesante del experimento es el vector beta. Le hemos llamado vector beta por simplicidad, pero en realidad es un vector de vectores, es decir, para cada característica virtual de cada concepto hay un vector beta de dimensiones n+1 siendo n el número de descriptores que agrupa la característica virtual.

En efecto, para cada concepto, generamos 19 vectores -uno por característica- que conjuntamente son los parámetros de la función que predicen el grado o la probabilidad de que una imagen tenga o no un determinado concepto.

Tal como vimos en el experimento, dada una imagen “I”, la decisión de clasificarla bajo un concepto “C” no es más que una función aplicada al vector de descriptores de la imagen I.

En primer lugar, aplicamos la función de distribución logística a cada característica virtual, siendo β_i los parámetros del vector beta de la característica virtual (i) correspondiente al concepto “C” en el que se está evaluando la imagen “I”, y X_i cada uno de los descriptores de la característica virtual (i) de la imagen “I”.

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \vec{\beta_i} * \vec{X_i})}}$$

Esta función nos devolverá un valor numérico entre cero y uno. La media ponderada de los 19 valores – uno por característica virtual – será la puntuación obtenida por esta imagen "I" para el concepto "C".

$$p = \frac{1}{k} \sum_{i=1}^{k} p_i \wedge k = 19$$

Ahora solo queda comparar este valor con el valor de la "equiprobabilidad" - que como vimos en el experimento no necesariamente ha de ser 0,5, sino que adoptamos la aproximación 1 que tomaba este valor entre la mínima puntuación de las imágenes positivas y la máxima puntuación de las imágenes negativas, en concreto su punto medio- y, si la iguala o supera, admitimos que la imagen "I" contiene el concepto "C", en caso contrario admitimos que la imagen "I" no contiene el concepto "C".

En definitiva es una función de decisión que necesita el vector beta del concepto "C", el límite de la equiprobabilidad para el concepto "C" y el vector de descriptores de la imagen "I". Esta función determina el veredicto -o sea, I es una instancia positiva o una instancia negativa de C- en función -valga la redundancia- de lo aprendido del concepto, que es el vector beta y el límite de su equiprobabilidad.

La imagen "I" instancia el concepto "C" si y sólo si:

$$\Longleftrightarrow F[Imagen(I), Concepto(C)] \geq equiprobabilidad[Concepto(C)]$$

En resumen, el vector beta es el elemento principal de lo aprendido del concepto "C". Este vector junto con el límite de la equiprobabilidad para el concepto "C" conforman el resultado del aprendizaje realizado en el experimento. La equiprobabilidad para un determinado concepto es un elemento significativo porque forma parte del aprendizaje, y además, puede utilizarse para reajustar rápidamente el concepto en los casos donde la predicción falle, sin necesidad de volver a realizar todo el proceso de aprendizaje, es decir, sin cambiar el vector beta.

En este sentido, una posible solución sería recalcular el valor de la equiproba-

bilidad que minimiza los errores tras la predicción fallida. Un pequeño algoritmo que podría implementarse como sigue:

> "*Si no hay errores, la equiprobabilidad la situamos en el punto intermedio entre la máxima de las puntuaciones negativas y la mínima de las positivas. Si no, recorremos todos las puntuaciones entre la mínima de las positivas y la máxima de las negativas para buscar en qué punto, si situasemos allí el corte, se cometen menos errores. Este punto será la nueva equiprobabilidad.*"

En notación algoritmica se puede ver en Algoritmo1, p. 51 y Algoritmo2, p. 52. El primero devuelve el valor de la equiprobabilidad tomando como entradas: 1) las puntuaciones de positivas y negativas entre la mínima de positivas y la máxima de negativas -intervalo crítico-, ambas incluidas; 2) el total de puntuaciones positivas o negativas del intervalo crítico; 3) puntuaciones de positivas del intervalo crítico, 4) puntuaciones de negativas del intervalo crítico; 5) puntuación mínima de positivas y 6) puntuación máxima de negativas. Se han desplegado todos los datos de entrada en aras a la claridad, esta entrada sería optimizable ya que hay redundancia de datos. Para obtener el resultado de salida, el Algoritmo 1 llama al Algoritmo 2 para que calcule el número de errores que se cometen tomando un determinado punto del intervalo crítico.

Algoritmo 1 Ajuste de equiprobabilidad

Entrada: L ← Puntuaciones de Positivas y Negativas entre la mínima positiva y la máxima negativa inclusivas; Errores ← Total de puntos entre la mínima positiva y la máxima negativa inclusivas; P ← Puntuaciones de Positivas entre la mínima positiva y la máxima negativa inclusivas; N ← Puntuaciones de Negativas entre la mínima positiva y la máxima negativa inclusivas; PM ← Puntuación mínima de Positivas; NM ← Puntuación máxima de Negativas.

Salida: Valor de la equiprobabilidad con menor tasa de errores.

```
L ← Puntuaciones de Positivas y Negativas entre la mínima positiva y la máxima negativa inclusivas
PM ← Puntuación mínima de Positivas
NM ← Puntuación máxima de Negativas
Errores ← Total de puntos entre la mínima positiva y la máxima negativa inclusivas
E ← Errores
si Errores = 0 entonces
  equipro ← (PM−NM)/2
si no
  para todo i ∈ L hacer
    erro ← llama_Cálculo_de_errores_según_corte(i, P, N)
    si erro < E entonces
      equipro ← i
      E ← erro
    fin si
  fin para
fin si
devolver equipro
```

Algoritmo 2 Cálculo de errores según corte

Entrada: Punto ← Punto de corte a estudiar; P ← Puntuaciones de Positivas y Negativas entre la mínima positiva y la máxima negativa inclusivas; N ← Puntuaciones de Negativas entre la mínima positiva y la máxima negativa inclusivas.

Salida: Valor de la equiprobabilidad con menor tasa de errores.

Punto ← Punto de corte a estudiar
P ← Puntuaciones de Positivas y Negativas entre la mínima positiva y la máxima negativa inclusivas
N ← Puntuaciones de Negativas entre la mínima positiva y la máxima negativa inclusivas
err ← 0
para todo i ∈ P **hacer**
  si $P_i <$ Punto **entonces**
    err ← err+1
  fin si
fin para
para todo i ∈ N **hacer**
  si $N_i \geq$ Punto **entonces**
    err ← err+1
  fin si
fin para
devolver err

2.3.6 ***El símbolo lingüístico.***

En este apartado solamente destacaré los posibles errores de etiquetaje y la posibilidad de utilizar el símbolo lingüístico como mediador para conectar con el procesamiento simbólico.

El etiquetado de la base de datos realizada por humanos se basa en la interpretación del símbolo lingüístico que representa al concepto. Además los etiquetadores residen en zonas con idiosincrasia diferente a la idiosincrasia de los proponentes de la clasificación y que además ni son expertos ni son homogéneos en sus decisiones. Por estas razones, el proceso está expuesto a un gran número de errores de etiquetado – y estos errores los entendemos a la vista de otro humano que en ningún modo le dotamos de la sabiduría suficiente para juzgar lo que el etiquetador ha tenido en cuenta o ha visto en la imagen en el momento concreto de su etiquetaje-.

Si además, este etiquetaje es la fuente sobre la que han de compararse los resultados del experimento, pocas esperanzas de arribar a una clasificación "verdadera" u objetiva independiente de los agentes humanos que la clasifican. El objetivo del experimento tampoco persigue categorizar bajo conceptos léxicos universales.

Aun así, este es el mundo real, y el experimento se ha realizado dentro de este contexto. Por tanto nos permitiremos dudar de la universalidad del símbolo bajo el cual caen las imágenes, pero admitiremos que en las relaciones humanas los conceptos suelen transmitirse como etiquetas léxicas y se supone un acuerdo en la interpretación de dichos símbolos, si es que el objetivo es entenderse en la comunicación.

2.4 *Interpretación de los resultados del experimento*

Con el análisis de las tasas de aciertos y errores de la fase 6 elaborado en la sección anterior (v. supra 2.2.3, p. 38), y teniendo en cuenta todos los datos recabados en el experimento incluidos en los anexos, vamos a intentar interpretar los resultados en relación a la línea de investigación desarrollada en esta parte del trabajo, que no es otra que *elaborar un constructor de hipótesis clasificatorias basado en regresiones logísticas cuyo objeto es clasificar imágenes bajo conceptos.*

Como toda teoría, las hipótesis elaboradas por este constructor han de ser validadas, pero no recurriremos a un falsacionismo estricto, à la Popper, para admitirlas o rechazarlas.[10] Nuestro margen de tolerancia con los fallos debe ser más amplio. Esta permisividad está justificada porque, al menos en el experimento realizado, la base de datos sobre la que se opera ha sido etiquetada por agentes humanos, sujetos a errores de apreciación subjetivos. Sólo por esta razón, ya cabe admitir un margen de tolerancia.

Por otra parte, ¿cómo explicar la inconsistencia entre la explicación y la predicción atendiendo a la aproximación utilizada en la equiprobabilidad? En parte eso queda resuelto por lo comentado en el apartado anterior: la desigual distribución de las imágenes positivas y negativas por cada concepto, que induce a maximizar el espacio probabilístico de los ejemplos mayoritarios -en este caso los negativos. Consecuentemente, será mejor realizar el corte lo más alto posible, es decir el 0,5 será preferible a los valores de corte más bajos obtenidos mediante la aproximación 1 (recuérdese, equiprobabilidad situada entre la mínima de las positivas y la máxima de las negativas, v. supra p. 42). En este sentido, podemos observar que incluso si este valor de corte se situara en 1, es decir, predecir que todas son negativas, las tasas de acierto serían efectivamente aún más altas. El riesgo, no obstante, sería ¡no acertar ninguna imagen positiva! Por eso, y con toda la cautela necesa-

[10]Debe matizarse un tanto el falsacionismo burdo ocasionalmente atribuido a Popper. No obstante, sus dificultades a la hora de dar cuenta de la contrastación de hipótesis probabilísticas (v. p. ej., Popper 1962, secc. 68) constatan las limitaciones de una metodología falsacionista en contextos donde la conexión entre hipótesis y evidencia no es deductiva. Nótese que dichas hipótesis, en principio y salvo que atribuyan valores probabilísticos absolutos, 0 o 1 –p.ej.: p (cara) = $\frac{1}{2}$– son lógicamente compatibles con cualquier frecuencia estadística empírica, o sea, finita, y por tanto no serían falsables, no serían científicas, según el criterio de demarcación defendido por el propio Popper.

ria, podríamos admitir la aproximación 1 también para las predicciones. Si así lo hacemos, el porcentaje de aciertos en imágenes positivas sube sustancialmente, a expensas de un descenso de aciertos en las negativas.

Puede pensarse que no debería tener el mismo peso el acierto/fallo en una imagen positiva que en una imagen negativa, dada la diferente proporción de imágenes positivas y negativas. Una sugerencia para una ponderación más adecuada sería puntuar aciertos y fallos con un peso inversamente proporcional a la frecuencia relativa con que aparecen en la base de datos. Cuanto mayor es la desproporción entre imágenes positivas y negativas, a favor de estas últimas, menor es el peso que cabe conceder a un acierto en una imagen negativa. De este modo, tendríamos una información más matizada para poder decidir qué aproximación es la mejor. Aunque este método puede ajustar estadísticamente la distribución de imágenes, nos encontrariamos con los problemas discutidos anteriormente en cuanto a la cantidad de imágenes positivas y negativas que debemos seleccionar para el entrenamiento y su repercusión en el valor de la equiprobabilidad (v. supra p. 42). Por otra parte, el contexto de uso condiciona qué aproximación debemos escoger y puede sugerir seguramente otros criterios adicionales. Por ejemplo, en una búsqueda médica de imágenes, un falso positivo tendrá una repercusión diferente a cuando buscamos un lugar barato donde pasar las vacaciones.[11]

Otro hecho importante es que, a pesar de elaborarse a partir de una base de datos etiquetada por humanos, *los resultados -explicativos y predictivos- son mejores cuando las imágenes de entrenamiento del modelo son elegidas automáticamente por una máquina frente a la elección manual realizada por un experto humano*. En principio, este último debería aportar mayor carga semántica, además de ser un agente del mismo tipo que quien clasificó las imágenes de la base de datos, con lo cual es natural suponer una mayor homogeneidad de criterios, salvando cierta idiosincrasia personal. De ahí que cabría esperar mejores resultados en las explicaciones y predicciones. En este sentido se puede comprobar que Baseline1, donde las imágenes positivas se eligen automáticamente a partir del centroide formado por

[11] Nótese también que en el primer contexto, a su vez, los fallos pueden tener una importancia bien distinta según sean positivos o negativos. Supongamos que se quiere averiguar si una imagen obtenida en un chequeo a un paciente instancia el concepto 'tumor con alta morbilidad'. Si la prueba arroja un falso positivo, el paciente recibirá un tratamiento que no necesita; pero si la prueba da un falso negativo, el paciente no recibirá el tratamiento, con el consiguiente riesgo de muerte.

las imágenes que instancian el concepto, obtiene un modelo que explica y predice mejor que Baseline2 y Baseline3, donde las imágenes positivas han sido elegidas por un experto humano. Este hecho se puede constatar en las tablas 2.18, p. 39; 2.19, p. 39, y en el anexo 6.2, p. 142, donde se recogen las configuraciones de las mejores explicaciones y predicciones.

Cabría pensar también la importancia de las familias. Parece que debería incrementarse el índice de acierto si las imágenes negativas se suministran dentro de la misma familia del concepto, ya que agudizamos el contraste entre imágenes pertenecientes a un universo más reducido, la propia familia. Este planteamiento es el de baseline 3, frente al de baseline 2 que suministra las imagenes negativas fuera de la familia. Según vimos en la sección anterior apartado cuarto (v. supra p. 42) el experimento no evidencia diferencias sustanciales en las tasas de aciertos/errores. Dado que el algoritmo construye su hipótesis a partir de las características puramente físicas de las imágenes, sin tener en cuenta la carga semántica de estas, estos resultados revelan que desde un punto de vista estrictamente físico no tiene por qué haber mayor similitud entre dos imágenes pertenecientes a dos conceptos de la misma familia que entre dos imágenes que pertenezcan a conceptos de distinta familia. Esto es plausible, en principio, y debería comprobarse calculando las distancias entre imágenes y centroides pertenecientes a diferentes niveles conceptuales (más particulares y más generales). Desde luego, la similitud aquí no se entiende en relación a los criterios de la percepción humana, sino en cuanto a las propiedades que selecciona la máquina.

En otro orden de cosas, es sugerente considerar las imágenes más próximas al centroide positivo como imágenes “prototipo”, y al centroide en sí como una imagen ideal cuya definición es puramente formal, lo cual entronca mejor con ciertas opciones teóricas sobre la estructura de los conceptos que serán comentadas en capítulos siguientes. Esta sugerencia hay que tomarla con cautela.

Si bien podemos considerar que, en la elección de imágenes en modo experto, el humano ha elegido para el conjunto de entrenamiento, las imágenes que mejor representan el concepto, es decir, ejemplos prístinos del concepto en cuestión, ejemplos que suscitarían un acuerdo si no universal, al menos muy amplio entre agentes humanos, presumiblemente imágenes prototipo, ¿podemos decir lo mismo en la selección automática?

En modo automático, la máquina elige como imágenes positivas de entrenamiento las más cercanas al centroide de todas las imágenes positivas, para el concepto que se está tratando, de toda la base de datos. El centroide y sus imágenes cercanas, por tanto, están fuertemente relacionadas con la distribución de las imágenes positivas en la base de datos, con la métrica utilizada y no con la posible semántica de estas.

Además, el experimento no ofrece evidencias de solapamiento entre la elección en modo experto y en modo máquina que nos permita atribuir la condición de imágenes prototipo a las imágenes elegidas automáticamente. En este sentido, si la efectividad mostrada por la elección automática es notablemente superior a la efectividad mostrada por la elección en modo experto, para nuestro experimento, ¿importa la capacidad representacional de las imágenes? Parece que no, al menos tal como entendemos la capacidad representacional desde la 'teoría prototípica de los conceptos' que veremos más adelante (v. infra 3.3, p. 87; 3.2, p. 78). Entendido así, el centroide positivo NO es un prototipo, es más bien una definición formal que podría ser satisfecha por muchas configuraciones de matiz, color, brillo y texturas, es decir, por muchas imágenes posibles, algunas de las cuales no tendrían ninguna capacidad representacional para un sujeto humano. Tampoco lo serían las imágenes próximas al centroide por su falta de representatividad según hemos discutido anteriormente.

Esta discusión sugiere dos vías alternativas:

- 1º) Si realmente la máquina es más efectiva que el humano en la tarea que nos ocupa (una tarea discriminativa), lo que esto indica es que dicha efectividad no requiere necesariamente introducir la variable de la capacidad representacional de las imágenes, su significado. Como modelo de comparación para discernir si una imagen, en virtud de su mayor o menor semejanza con el prototipo, ejemplifica o no el concepto no se requiere, pues, captar o almacenar ninguna imagen con contenido representacional (como es el "prototipo", en el sentido técnico que esta palabra tiene en la teoría prototípica sobre los conceptos).

- 2º) Redefinir el concepto teórico de prototipo para la máquina. Muy diferente al que sostienen las teorías prototípicas de conceptos. Cuyo sentido

recaiga sobre la métrica de los descriptores y no sobre el significado de las características (v. infra 3.2, p. 84). En cuyo caso, el prototipo quedaría completamente desprovisto de capacidad representacional, y ese es justamente su valor principal en relación a cómo los agentes humanos operan con conceptos, según las teorías mencionadas.

Por otra parte, es claro que, aunque tengamos una hipótesis que explique bien las imágenes de prueba eso no significa que sea capaz de predecir con la misma efectividad. Y lo que realmente nos interesa aquí es la capacidad predictiva, ya que lo que tratamos de averiguar es si la máquina adquiere/aprende un concepto, y un aspecto importante relacionado con eso es la capacidad de discriminar de modo efectivo las instancias positivas de las negativas, pero no las que ya se conocen, claro (las que estaban incluidas en la base de entrenamiento), sino las que aún no se conocen, de ahí que el factor más relevante para nosotros sea la efectividad predictiva.

Respecto al mismo concepto podría haber variaciones significativas en el rendimiento, al alza o a la baja, cambiando una base de datos por otra. Apenas tiene sentido referirse a la base de datos 'Verdadera', con mayúsculas. En todo caso, siempre necesitaremos una base de datos que se habrá etiquetado/clasificado por agentes humanos; con mayor o menor cuidado, pero por agentes humanos. Lo que se desprende del experimento expuesto es:

- *primero*, que la clasificación inicial de una base de datos es imprescindible;
- *segundo*, que no la realiza la máquina y por tanto se hace siguiendo criterios semánticos (capacidad representacional de las imágenes);
- *tercero*, que lo que la máquina hace es, a partir de dicha clasificación semántica, generar un criterio de clasificación formal de las imágenes que aspira a igualar o mejorar el rendimiento de los agentes humanos en tareas discriminativas;
- *cuarto*, que la máquina no intenta simular los procedimientos efectivamente seguidos por un clasificador humano; ese no es el objetivo aquí, ya que no se trata de simular el comportamiento de los agentes humanos, sino elaborar un criterio alternativo y comprobar su potencia explicativo-predictiva.

Entonces, ¿cuántos fallos serían necesarios para considerar que no se ha aprehendido el concepto en cuestión? ¿Podría el ordenador cometer un fallo clamoroso, como podría ocurrirle a un sujeto humano como consecuencia del cansancio o la distracción momentánea? No parece que tenga mucho sentido estipular un porcentaje concreto de aciertos/errores para determinar el grado de adquisición de un concepto por parte del ordenador. Ahora bien, aunque eso sea así, la eficacia puede plantearse en términos comparativos con lo que hace el humano (v. infra anexo 6.5, p. 146). El agente humano comete fallos debido al cansancio, etc., que la máquina no comete. Por eso, una tasa de rendimiento igual entre la máquina y los agentes humanos no sería un resultado muy satisfactorio para la primera, ya que descontando aquellos errores humanos, la máquina sería menos eficaz. No obstante, para atribuir a la máquina cierta capacidad de aprendizaje no haría falta exigir un rendimiento superior al humano. Téngase en cuenta que la máquina comienza desde cero, por así decirlo, y tras realizar el proceso descrito en el experimento, es capaz de igualar e incluso superar al agente humano. Algún aprendizaje se ha producido, si no, no podríamos explicar los resultados.

Por otro lado, una tasa de aciertos/errores similar entre el humano y la máquina no significa que esté resuelta la brecha semántica (v supra sec. 2.2, párrafo 3, p. 15). Aunque estadísticamente ambos sean iguales, tenemos evidencia que, ni los errores ni los aciertos, cometidos en la clasificación por humanos y por la máquina coinciden. En este caso podemos afirmar que si bien los errores de los humanos podemos considerarlos como errores cometidos por fallos semánticos de interpretación (una vez descartados los errores de cansancio, etc.), en el caso de la máquina (que en la ejecución de su proceso no ha cometido fallos) no pueden ser atribuidos a ninguna causa semántica, sino a la función de predicción sobre los descriptores de bajo nivel. En resumen, que ambos procesos muestren una eficiencia estadística similar, significa que son igual de eficaces para la tarea de clasificación, pero no supone la desaparición de la brecha semántica entre los descriptores físicos de la imagen y la interpretación semántica que de la misma imagen se forma un humano.

En todo caso, a la presente discusión subyace la idea de que aprender o adquirir un concepto está íntimamente relacionado con una habilidad discriminativa clasificatoria, por así decirlo. En los siguientes capítulos se exploran justamente estas cuestiones.

Ningún conocimiento humano puede ir más allá de su experiencia.

John Locke

CAPÍTULO

3

Discusión del experimento en relación al debate contemporáneo sobre la naturaleza de los conceptos

El capítulo 2 de esta tesis mostró un experimento de clasificación en inteligencia artificial, con el fin de arrancar la discusión sobre el tema principal de la tesis, ***si las máquinas pueden poseer conceptos***. En este sentido, los resultados sugerían abrir una ventana a la discusión de las principales teorías filosóficas y de las ciencias cognitivas centradas sobre la máquina. Para poder avanzar en este camino, la sección 2.3, p. 43, de esta tesis identificó los elementos más significativos del experimento susceptibles de ser discutidos en base a estas teorías.

Ahora, en esta parte de la tesis abordaremos la discusión filosófica sobre la pertinencia de la hipótesis principal, y lo haremos discutiendo y tomando partido sobre el encaje de los principales elementos del experimento expuestos en la sección 2.3, p. 43, dentro de las teorías conceptuales de la filosofía y las ciencias cognitivas.

En primer lugar abordaremos la problemática de identificar lo conceptual y lo

no conceptual dentro de los procesos de pensamiento de las distintas teorías cognitivas, buscando el encuadre de los elementos más significativos del experimento. En segundo lugar se intentará enmarcar los resultados del experimento y los elementos significativos dentro de las principales teorías de conceptos. En tercer lugar se discutirá la naturaleza de los conceptos, su estatus ontológico, y se tratará de ubicar los posibles proto-conceptos identificados en el experimento. Y por último justificaremos que los posibles conceptos aprendidos por la máquina no son innatos.

3.1 *Lo conceptual y lo no conceptual*

Desde el punto de vista del aprendizaje humano, parece que hay un amplio acuerdo entre los filósofos y los psicólogos actuales en que los conceptos son constituyentes básicos del pensamiento humano, elementos fundamentales para realizar los procesos de categorización, inferencia, memoria, aprendizaje y toma de decisiones. Es importante reseñar, además, que el marco genérico de las investigaciones en psicología a menudo se ha inspirado en fuentes filosóficas, y en este sentido cabe destacar hitos filosóficos como la distinción de Frege entre sentido y referencia, la noción de "aire de familia" de Wittgenstein, la tesis de que la única acepción legítima de significado es el "significado estimulativo", defendida por Quine, o las discusiones sobre externismo –los significados no están "en la mente"- y esencialismo de Kripke y Putnam.

Aún así, la pregunta "¿Qué es un concepto?" sigue siendo un tema espinoso. Los teóricos discrepan sobre "¿Qué son los conceptos? ", "¿Qué tipos de fenómenos explican?" e incluso se plantean si realmente existen los conceptos. Algunos como Fodor (Fodor 1998; Margolis & Laurence 2007) toman los conceptos como representaciones mentales significativas que se combinan para formar pensamientos enteros, en cambio otros como Peacocke (Peacocke 1992; Zalta 2001) los toman como entidades abstractas que se componen para formar los contenidos proposicionales que expresan los pensamientos.

En este apartado intentaré distinguir lo conceptual de lo no conceptual para aplicarlo a los resultados del experimento expuesto en el capítulo 2, y, en concreto sobre los elementos recopilados en la sección 2.3, p. 43. Para este fin, me apoyaré en

la discusión que plantean Margolis y Laurence en su artículo sobre el alcance de lo conceptual (Margolis & Laurence 2012).

El primer obstáculo que nos encontramos está en que la mayoría de las investigaciones filosóficas y de las ciencias cognitivas que tratan esta discusión centran su dominio sobre el humano adulto,- con algunas excepciones que afectan a niños en estado pre-léxico y algunos animales,- aunque este hecho no debe impedirnos abordar un análisis crítico en el que pueda expandirse al dominio a los autómatas.

En este sentido, debemos identificar, si es posible, a qué elementos de los autómatas podemos hacer referencia cuando en las discusiones filosóficas se mencionan los conceptos como construcciones de la mente humana.

En primer lugar, los estados mentales de los que discuten la mayoría de los argumentos filosóficos pueden parecer difíciles de trasladar a un autómata. Aunque, Michael Dummet (1993a,1993b; citado en Margolis & Laurence 2012) argumenta que los animales no son capaces de pensamiento conceptual de pleno derecho, sino que tan solo exhiben una disminuida forma de pensamiento que llama proto-pensamiento, se podría trasladar esta idea sobre las máquinas y hablar de una nueva construcción para identificar la mente de la máquina, una "proto-mente" y admitir así estados proto-mentales en las máquinas. Además, sería plausible definir los estados de creencias en los autómatas como aquellos estados cuyo contenido ha sido aprendido, no diferenciándose, como concepto, excesivamente de los estados de creencias de los humanos. En la misma dirección podemos definir los estados perceptuales de los autómatas como aquellos cuyo contenido es el resultado de la captura de la realidad empírica por medio de aparatos mecánicos simulando órganos sensoriales.

En segundo lugar, las diferentes clases de estados mentales en humanos difieren en muchos aspectos, según los teóricos, pero no debemos asumir que deban ser divididos en estados conceptuales y estados no conceptuales, al menos en principio, ya que la distinción conceptual / no conceptual puede ser descrita en función de los tipos de contenido o sobre diferentes tipos de estados representacionales.

Si nos basamos en el tipo de contenido para distinguir lo conceptual de lo no conceptual, este lleva implícito la existencia de contenido no conceptual, de forma que admitiremos el supuesto de que existen estados mentales y que son portadores de contenido conceptual y/o no conceptual, además, la tipificación de contenido

coneptual/no conceptual a menudo depende del modelo de concepto. Una visión minimalista y una visión más rica diferirán en qué debe considerarse contenido no conceptual frente al conceptual. (Bermúdez & Cahen 2015).

Por otra parte, un punto de consenso entre los filósofos que alimentan estos debates es que los constituyentes de las representaciones o contenidos que están involucrados en los estados de creencias paradigmáticas deben contar como conceptos. Estos estados incluyen las creencias conscientes que tiene un humano adulto que no haya sufrido daño cerebral o anormalidades que provoquen alteraciones cognitivas o lingüísticas.

Por otra parte, en (Margolis & Laurence 2012) se revisan varios argumentos que pretenden mostrar que algunos estados perceptivos son no conceptuales, pero ninguno de ellos es concluyente en este sentido, y aún más, si pretendemos incluir a la máquina dentro del dominio de aplicación de estos argumentos, deberíamos revisarlos con más detenimiento.

Así, el argumento de *la finura de grano*, que afirma que el estado perceptivo soporta capacidades de discriminación que son considerablemente de más grano-fino que nuestra inventiva de conceptos (Evans 1982; Peacocke 1992) , puede ser adaptado fácilmente si admitimos que en nuestro experimento, la imagen bidimensional se representa en el espacio RGB o HSV como una matriz numérica de tres columnas cuyas dos primeras columnas corresponden con la posición del pixel y el tercero con cada uno de los canales, la resolución de la imagen depende de la cantidad de bits asignados para su profundidad, siendo usual utilizar un byte (ocho bits), en cuyo caso los valores oscilan entre 0 y 255, mientras que tras las transformaciones de cuantización, esta imagen queda representada por un conjunto numérico menor, que es de donde arranca el proceso de clasificación en el experimento. La propuesta es que un espacio de dimensiones menores que representa a uno de dimensiones mayores implica una conceptualización en el sentido que hay una agrupación de instancias del estado de más dimensiones (perceptivo) que caen bajo una misma instancia en el estado de menos dimensiones (de los 'conceptos').

Pero esto no resuelve el problema, ya que también podrían contener conceptos los estados perceptivos. Y ciertamente, en el experimento, la codificación de las imágenes en base a sus descriptores, es decir tras el proceso de cuantización, cumple con las exigencias del argumento pero no podemos considerar que estemos ante

estados de creencias, aún no hemos aprendido nada, aunque podamos admitir un cierto proceso de conceptualización en este estado de la percepción que es capaz de pre-conceptualizar la imagen. Desde esta perspectiva, se podría admitir que los estados perceptivos, además de contenido no conceptual como la matriz RGB o HSV de una imagen, contienen una especie de contenido pre-conceptual representado por el punto en el hiper-espacio multidimensional de los descriptores de imagen. Es decir, la imagen queda representada por su vector de descriptores.

En cierta medida, este argumento está relacionado con el de *la riqueza de la experiencia*, que plantea que es mucho más rica en detalles la experiencia perceptiva que la conceptual, tanto que puede desbordar los recursos del sistema conceptual. Peacocke en (Peacocke 2001) argumenta que el contenido que tiene el estado perceptivo es de clase diferente al contenido asociado con el estado de creencias. Para sostener este argumento Peacocke propone un escenario orientado y centrado desde la posición del que percibe. En este escenario, y en lo concerniente a la visión, traza ejes espaciales y asigna a cada punto, si es una superficie, textura, matiz, brillo y saturación[12]. Aclara que la imagen se especifica así con gran detalle, y que difiere completamente de los contenidos de las creencias en términos de las tramas semánticas que conforman su composición. Con otras palabras, mientras los contenidos conceptuales están asociados con tramas semánticas proposicionales tipo sujeto-predicado, los contenidos no conceptuales están asociados con tramas semánticas no proposicionales[13] que nos dan directamente el escenario. Parece que la distinción conceptual / no conceptual esté centrada en el tipo de trama semántica asociada, si es proposicional o no proposicional, pero esta discusión no está terminada, pues también es posible expresar los estados perceptivos en términos de tramas semánticas proposicionales, más complicadas eso sí, para poder reflejar los detalles de la percepción[14].

[12]Nótese que justamente estos parámetros son los que utiliza nuestro experimento para codificar la imagen digital.

[13]En este sentido, una trama semántica proposicional es aquella que puede expresarse en el lenguaje de la lógica proposicional, mientras que una trama semántica no proposicional, no. Véase la diferencia entre la expresión 'una casa verde' que seria reductible a una trama semántica proposicional y una fotografía de 'una casa verde' con todos sus matices de color y textura.

[14]Nótese que podemos expresar mediante el lenguaje -semántica proposicional- la descripción de una imagen visual. En este sentido, podemos enunciar con dicho lenguaje incluso los detalles más finos de la percepción, eso sí, cuanta más precisión necesitemos para describir la imagen, más complicada será la trama semántica proposicional asociada, en cambio, la imagen codificada en una

3. DISCUSIÓN DEL EXPERIMENTO EN RELACIÓN AL DEBATE CONTEMPORÁNEO SOBRE LA NATURALEZA DE LOS CONCEPTOS

Para nuestro interés, tomaremos partido por la propuesta de Peacocke, es decir, hay contenidos no conceptuales diferentes de los contenidos conceptuales que se encuentran ligados a estados de percepción, y que, en el caso de la visión bidimensional, concuerdan con la matriz de colorimetría en el espacio HSV; además, en el experimento también hemos extraído descriptores de textura y de forma, todos ellos candidatos a ser contenido no conceptual similar a la propuesta que Peacocke plantea para los humanos. Aunque tomemos la matriz HSV como no conceptual y ligada a los estados de percepción, todavía no podemos decir nada del vector de descriptores que representa a la imagen. Lo que sí parece claro es que ni la matriz HSV ni el vector de descriptores están ligados a tramas semánticas proposicionales, y por tanto no deben ser considerados como conceptos.

En la misma dirección del argumento anterior, que las representaciones perceptivas son no conceptuales, Fodor (Fodor 2007 y 2008) propone otra vía para caracterizar la distinción conceptual / no conceptual, y la expresa en términos del *contraste de lo que llama formas discursiva e icónica de representación*. La clave diferencial está en cómo representan las partes de un todo. El lenguaje es un paradigma de la forma discursiva, en este sentido las oraciones admiten una descomposición canónica donde se pueden identificar los conceptos, por ejemplo "Juan come pan" admite descomposición canónica del tipo ("Juan", "come", "pan"), mientras una fotografía, que es un ejemplo de forma icónica, no admite esta descomposición y propone para estas el principio del dibujo, que dice que: *"Si P es un dibujo de X, las partes de P son dibujos de las partes de X"*, principio que no es satisfecho por la forma discursiva de representación. En definitiva afirma que las creencias, como las oraciones, son discursivas y por tanto conceptuales, mientras que los estados perceptuales, igual que las fotografías, son no conceptuales. Además afirma que sólo los conceptos implican representaciones en el sentido de *"representado por"*.

Y si admitimos la última afirmación, ¿podría el vector beta ser candidato a concepto? El vector beta son los coeficientes que hay que aplicar multiplicando a los descriptores para que su suma final, tras pasar por una función sigmoide, nos indique la probabilidad de que una determinada imagen instancie un determinado concepto, teniendo en cuenta el punto de equiprobabilidad (para mayor detalle v. supra capítulo 2, sección 2.3.5, p. 48). Si tomamos el lenguaje matemático, podríamos

trama semántica no proposicional no cambia su nivel de complejidad, es siempre la misma imagen.

definir la función:

Instanciar concepto: F(CONCEPTO(C), Imagen(I)) $\geq$ límite de la equiprobabilidad (C).

De tal manera que toda imagen que cumpla esta función, por el hecho de instanciar el concepto C, será una representación (más o menos precisa) de dicho concepto C, y podemos decir que el concepto (C) está representado (no de manera exclusiva) por la imagen (I).

Dada esta definición, la función F se convierte en parte de un lenguaje con sujeto y predicado, y además respeta que los conceptos impliquen representaciones de imágenes, es decir muchas imágenes que instancian un mismo concepto, están representadas por este. Además, una imagen que instancia un concepto se convierte en una representación, más o menos precisa, de este concepto en el escenario discursivo donde aparezca la imagen. En definitiva, podemos afirmar la construcción en lenguaje que dice:

" La Imagen (I) 'representa a' un CONCEPTO (C)" si y sólo si F(CONCEPTO(C), Imagen(I)) $\geq$ límite de la equiprobabilidad (C).

Matizando que, aunque en general, la relación representacional no es simétrica, es decir, si un concepto representa a una serie de instancias particulares (imágenes en nuestro caso), no es lo mismo que afirmar que cada una de estas imágenes representa al concepto con todo su contenido. Sí podemos admitir que la imagen que instancia un concepto puede ser candidata a representar dicho concepto al menos parcialmente, en el contexto donde se esté utilizando. De esta forma, una imagen de un día soleado, aunque no sea la única que instancie el concepto 'timeofday_day', este concepto podría estar representado por esta imagen.

Para nuestro interés, la diferencia entre el argumento defendido por Fodor frente al defendido por Peacocke radica en que mientras este último asigna el contenido vehicular de la imagen como una matriz numérica donde se codifica la posición y las características de color, textura y forma de cada punto, al contenido no conceptual, Fodor plantea su forma icónica como la imagen misma, sin prestar atención al vehículo. Lo que queda claro es que ambas parecen candidatas a ser contenidos perceptivos y por tanto contenidos no conceptuales. Y ambas formas de representación se pueden identificar fácilmente en el experimento realizado. Volvemos a considerar la matriz HSV y quizá también el vector de descriptores que representa

a la imagen como no conceptual, y además, con Fodor podemos extender lo no conceptual a ***las imágenes mismas***, es decir, su representación en pantalla, no su matriz numérica.

Hasta ahora hemos prestado más atención en identificar contenidos no conceptuales, pero no hemos avanzado mucho en identificar los contenidos conceptuales. En primer lugar mi interés en localizar lo no conceptual en la mente de los humanos aspira a encontrar sus homólogos -si los hay, y parece que sí- en nuestro experimento. Realizada esta tarea, algunos de los contenidos conceptuales de la mente humana se han construido a partir de contenidos no conceptuales, y este será otro espacio de nuestro interés.

En el artículo mencionado anteriormente (Margolis & Laurence 2012) , los autores recopilan varios argumentos para dar soporte a la idea de que los animales y los niños sólo poseen estados no conceptuales, - o que carecen de conocimiento genuino- y por tanto son incapaces de poseer conceptos. Y si es así, ¿están confinados a representar el mundo usando estados mentales significativamente diferentes de los estados mentales que disfruta un adulto? Parece que sí. Pero esto no significa que no puedan representar el mundo, a su manera, y si lo representan ¿acaso no necesitan conceptos – diferentes a los que tienen los humanos adultos, eso sí - para representarlo?

Si podemos admitir la existencia de conceptos diferentes a los conceptos que poseen los humanos adultos, en base a observar comportamientos que necesitan de un conocimiento específico y que este conocimiento esté codificado en conceptos con los cuales se ha producido el pensamiento que ha llevado a la acción, ***¡qué más da quién sea el autor, adulto humano, niño, animal, o máquina!*** Lo que sí importa es cuáles son los criterios que definen la posesión de un concepto.

Aún nos queda esta última cuestión por tratar, ¿quién tiene conceptos? Si analizamos los argumentos filosóficos que se aportan, en el citado artículo de Margolis y en (Bermúdez & Cahen 2015) , para identificar "quién tiene conceptos", podemos determinar "qué tipo de conceptos tiene", y plantear, si fuera necesario, otro tipo de conceptos adecuados para la máquina. Curiosamente, casi todos basan directa o indirectamente esta capacidad en el lenguaje humano.

En este sentido, Donald Davidson, en el *argumento de la opacidad*, afirma que el lenguaje es necesario para las representaciones, ya que es el que permite inter-

pretar el discurso (Davidson 1975). Nos invita a pensar que la distinción conceptual / no conceptual se basa en la capacidad de sostener representaciones en el sentido de representaciones distintas de un mismo referente en diferentes discursos. Algo parecido hemos visto anteriormente en el argumento de Fodor, que afirma que las creencias, como las oraciones, son discursivas y por tanto conceptuales, además, reclama que la capacidad de un concepto para poseer diferentes representaciones, sólo se concibe si hay un lenguaje. Para Davidson sólo los humanos son capaces de tener representaciones diferentes de un mismo referente en diferentes discursos. Pero este argumento cae tras probar que otras especies pueden también distinguir distintas representaciones de un mismo concepto referente, por ejemplo, en los Babuinos, el hijo de un jefe ve a su padre como padre o como jefe dependiendo de las circunstancias discursivas del entorno (Bergman et al. 2003).

Para este tipo de contenidos, ya hemos discutido en el argumento de Fodor sobre forma discursiva y forma icónica (v. supra p. 66), que el conjunto: ***vector beta, función logit, límite de equiprobabilidad***, cuando entran en contacto con los descriptores de una imagen pueden expresarse en un contexto discursivo (en concreto un lenguaje lógico), si consideramos que una imagen (I) instancia un concepto (C), y que el concepto está representado por el vector beta (B), podemos aceptar que distintas imágenes (I) instancian un mismo concepto (C). Si como allí discutimos, la imagen (I) que instancia un concepto (C) representa, al menos parcialmente, a este, entonces, admitimos distintas representaciones (imágenes I) de un mismo referente (concepto C). Pero no hemos cambiado de contexto, son distintas representaciones dentro del mismo contexto, es decir, estamos identificando concepto (C) con vector beta (B). También podemos cambiar el conjunto de imágenes de entrenamiento para un nuevo contexto, del mismo concepto por supuesto, y obtener así un nuevo vector beta (B') y un nuevo límite de equiprobabilidad, ¿realmente podemos considerar a ambos (B, B') como un mismo concepto referente (C)? Entendemos que no, básicamente por dos razones. Primera, para poder sostener estas dos representaciones (B, B') dentro de un mismo concepto referente (C) necesitamos una red externa que sea capaz de determinar en qué contexto específico ha de utilizarse cada una de las representaciones (B, B'), y este aspecto no ha sido tratado en el experimento. Y segundo, tal como se ha enfocado el experimento, el resultado de volver a generar un nuevo vector beta y su límite de equiprobabilidad apuntaría a

un concepto referente nuevo (C'), y no a una representación distinta de un mismo referente (B', C).

Davidson también argumenta que para tener creencias, y por tanto conceptos, hay que tener creencias sobre creencias, y esto solo es posible si se tiene un lenguaje natural. Es el *argumento metacognitivo*. Afirma que para modificar una creencia hay que tener el concepto de verdad y falsedad y estos conceptos son del lenguaje, ya que emergen del discurso. Por tanto concluye que para tener creencias se necesita el lenguaje.

En nuestro caso admitimos que no es necesario tener creencias para tener conceptos, es decir, admitimos la posible existencia de conceptos en la máquina sin necesidad de que esta tenga creencias.

Otro argumento de Davidson sobre la posesión de conceptos, *el argumento del holismo*, apunta a la necesidad de tener una rica red de inferencias para los contenidos conceptuales, de forma que el significado de un concepto esté determinado por la red de inferencias (Davidson 1975, 2004). Con este argumento es difícil negar contenido conceptual a algunos animales observando su comportamiento, aunque podamos admitir que un perro no tenga el mismo concepto de árbol que un humano, ya que el humano tiene una red de inferencias quizá más rica, no podemos concluir que el perro no tenga ningún concepto de árbol. El objetivo perseguido por nuestro experimento es más modesto. Solamente pretende encontrar el vector beta de un posible concepto y fijar el límite de equiprobabilidad -aprender el concepto-. Nada dice de tener redes de inferencia para los contenidos conceptuales. Por tanto, si admitimos que los conceptos no se poseen aisladamente sino en redes de conexiones inferenciales, tal como afirma el argumento del holismo, nuestro experimento no evidencia posesión de conceptos.

Otro argumento relevante respecto a la atribución de conceptos fue planteado por Gareth Evans. Según este autor, los que poseen conceptos pueden generar un número indefinido de nuevos pensamientos desde los conceptos que ya poseen. Estos nuevos pensamientos obedecen a lo que Evans llamó "condición de generalidad" [*generality constraint*]. Como dice Evans:

> "..if a subject can be credited with the thought that a is F, then he must have the conceptual resources for entertaining the thought that a is G, for every property of being G of which he has a conception". (Evans

1982, p. 104).

> "We cannot avoid thinking of a thought about an individual object x, to the effect that it is F, as the exercise of two separable capacities; one being the capacity to think of x, which could be equally exercised in thoughts about x to the effect that it is G or H; and the other being a conception of what it is to be F, which could be equally exercised in thoughts about other individuals, to the effect that they are F". (Evans 1982, p. 75).

Con otras palabras, se les exigen dos capacidades. Por una parte el pensar en el objeto individual 'x', que le permite relacionarlo con distintas clases ('x' es F y 'x' es G). Por otra parte, pensar en la clase de forma que se pueda predicar de otro objeto individual 'z' que es F. Es decir, se exige recombinabilidad sistemática.

Por ejemplo, si tenemos los conceptos 'feo' (F) y 'guapo' (G) y los individuos Juan (j) y Miguel (m), hemos de ser capaces de pensar 'Fj' y también 'Fm'. Para poseer el concepto F, debemos poder recombinarlo con todos los individuos. "Juan es feo" y "Miguel es feo". Si no podemos tener el pensamiento "Juan es feo" o "Miguel es feo", nos falla la recombinabilidad, y según Evans, el sujeto no tendría el concepto 'feo'. Además, debemos poder formar todos los conceptos 'Fj', 'Gj', 'Fm' y 'Gm'. Pero el requisito exigido a la posesión de conceptos es demasiado fuerte para algunos autores que abogan por una gradación en la posesión de conceptos. (Camp 2009; Beck 2012). Relajar el requisito de recombinabilidad sistemática es una opción para dar cabida a contenido conceptual en criaturas pre-lingüísticas.

En este sentido, muchos de los estudios relacionados con los conceptos parten del estudio de la conducta de animales y humanos. Pretenden a partir del comportamiento frente a estímulos externos, determinar si sus respuestas requieren algún tipo de pensamiento. En concreto si puede derivarse de este comportamiento que ha sido necesario un pensamiento conceptual. A partir del reconocimiento de cierto pensamiento conceptual, se intenta determinar que concepto de 'concepto' se requiere para soportar los resultados obtenidos.

En esta linea el artículo de Elisabeth Camp (Camp 2009) trata de encontrar diferentes grados de definición de 'concepto'. En concreto propone la definición de tres tipos de nociones de 'concepto'. Primero, una noción *"minimalista"* que de-

nota habilidades cognitivas de representación que son causalmente recombinables en un escenario contrafáctico. Segundo, una noción *"moderada"* que requiere habilidades cognitivas representacionales que se pueden recombinar sistemáticamente, independientes del estímulo exterior recibido y autogeneradas de forma activa. Tercero, una noción *"intelectualista"* que además de las propiedades de la noción moderada se le añade la capacidad de reflexionar, donde supone que esta capacidad solo es posible en el contexto de un lenguaje. En este sentido, Camp no busca una definición única de concepto, ni discutir cuál de ellos es teóricamente más útil, sino una gradación de estos, que por supuesto pueden convivir dentro de un mismo ente.

Afirma que el concepto "minimalista" tiene la ventaja de la parsimonia, ya que está motivado por las tareas básicas de los conceptos -representación y razonamiento sobre datos-, y permitiría responder de forma inteligente ante una amplia gama de estímulos con capacidades cognitivas relativamente modestas. Pero este concepto minimalista tiene un compromiso con el mundo pasivo, no se le ve comprensión activa, con otras palabras, responde al esquema estímulo-respuesta de forma pasiva.

Para tener pensamientos de forma activa es necesario el concepto "moderado", que no necesita ningún estímulo para arrancar, sostiene que a mayor independencia del estímulo, mayor recombinabilidad con lo que potencia el razonamiento instrumental -habrá más pensamientos sobre qué puedo hacer con algo-, además las capacidades cognitivas del ente se vuelven más flexibles. Juntas, la capacidad de recombinabilidad sistemática y la independencia del estímulo potencian el objetivo más básico del pensamiento: el uso de la información sobre el mundo para resolver problemas y facilitar al ente su supervivencia. Finalmente, para el concepto "intelectualista", afirma que el lenguaje facilita notablemente la independencia del estímulo y la recombinabilidad, además permite "escuchar" los pensamientos de los demás y posibilita poder retomar el mismo pensamiento en diferentes situaciones. El mismo lenguaje tiene sus propias reglas sintácticas para incrementar la recombinabilidad, y si es suficientemente potente, tendrá medios para denotar valores de verdad y representaciones inferenciales entre distintos pensamientos.

En nuestro experimento no se cumple la 'generality constraint'. No se puede hablar de recombinabilidad sistemática. Aunque puedan darse casos particulares que sí la cumplan. Por ejemplo:

- Sean Imagen1 ($x1$), Imagen2 ($x2$) instancias de imágenes representadas por

sus respectivos descriptores.

- Sean Concepto1 ($C1$), Concepto2 ($C2$) instancias de conceptos representados por sus centroides, sus vectores beta y sus límites de equiprobabilidad.

- Sean C1 'timeofday_day' y C2 'view_portrait'.

La 'generality constraint' impone que se puedan construir todas las combinaciones posibles.

Podría darse el caso de $x1C1$ y $x2C1$, que ambas imágenes instanciaran el concepto 'timeofday_day', y al mismo tiempo $x1C2$ y $x2C2$ que ambas imágenes instanciaran también el concepto 'view_portrait'. Sería el caso que $x1$ y $x2$ sean imágenes de personas tomadas de cerca, al aire libre, un día claro. Aquí sí se cumple la recombinación sistemática, como un caso particular.

En general, para todas las instancias x_i y para todos los conceptos C_i, no se va a cumplir.

Si somos estrictos en la aplicación de la 'generality constraint' no podemos considerar como contenido conceptual las representaciones de la máquina de nuestro experimento. A esta misma conclusión llega Beck (Beck 2012) cuando afirma que las representaciones primitivas de magnitudes espaciales, temporales, numéricas y otras derivadas de la percepción, a las que él llama 'magnitudes analógicas', son no conceptuales por carecer de la necesaria recombinabilidad que exige la 'Generality Constraint' de Evans.

Nuestra experimento muestra una máquina que ante la entrada de una imagen, reconoce una serie de conceptos. ¿Hay razonamiento sobre datos en este proceso? Entendemos que sí. Primero, en la tarea de aprendizaje, cuando se calcula el vector beta a partir de los descriptores de las imágenes de entrenamiento hay un razonamiento sobre datos (los descriptores) intrínseco al cálculo de los parámetros del vector beta y el límite de equiprobabilidad. Segundo, ante una entrada (descriptores de imagen) hay otro cálculo con el vector beta y el límite de la equiprobabilidad para razonar si esta entrada instancia un determinado concepto. Por otra parte, y tal como se comentó en el párrafo anterior, no se cumple la recombinación sistemáticamente como exige la 'restricción a la generalidad' de Evans, pero sí podemos ver escenarios donde causalmente se muestre recombinabilidad. Son los casos en los

que ante una imagen de entrada esta instancie más de un concepto. En estos casos se puede afirmar que la imagen (I) es el concepto (C) y también el (C') tal como exige la recombinabilidad. Tampoco se detecta que los procesos puedan actuar con independencia de estímulo externo, es decir, para iniciar el proceso se debe presentar ante la entrada una imagen, y a partir de ahí la máquina actúa, por tanto no son aplicables ni el concepto moderado ni el concepto intelectualista de Camp que exigen independencia de estímulo externo. En resumen, el hecho de que la máquina muestre unas habilidades cognitivas como la representación y la discriminación, que se aprecie un mínimo razonamiento sobre datos, además que pueda causalmente recombinar las representaciones (con las restricciones comentadas) nos acerca, al concepto 'minimalista' de Camp (Camp 2009).

Finalmente, consideremos por un momento, que todos los contenidos de nuestro experimento son contenidos perceptuales. Si aceptamos que los contenidos de la percepción son no conceptuales. Si además aceptamos que existen conceptos observacionales derivados de la percepción. Entonces, estamos admitiendo que existe un proceso de aprendizaje de conceptos derivados de la percepción que parte de contenidos no conceptuales a contenidos conceptuales. Si no se admite este proceso de aprendizaje, estaríamos comprometiéndonos con un innatismo radical (v. infra 3.4, p. 96), es decir, los conceptos observacionales derivados de la percepción, por ejemplo un 'árbol', son innatos. Y esta postura es difícil de defender como sostiene Roskies en (Roskies 2008), por ser incompatible con las principales teorías de aprendizaje de conceptos.

Según este argumento, en nuestro experimento, los contenidos que no derivan de ningún proceso de aprendizaje, serían no conceptuales. Pero, tanto el vector beta como el límite de equiprobabilidad, que derivan de un proceso de aprendizaje, y para mantener el argumento de Roskies, no pueden considerarse 'no conceptuales'. ¿Es suficiente este argumento para considerar como contenidos conceptuales al vector beta y al límite de equiprobabilidad? Y si es así ¿no deben ser conceptos todos los contenidos conceptuales? Bien, quizá no sean conceptos como los conceptos observacionales que se le atribuyen a un humano, 'árbol', 'casa', etc. entre otras razones porque estos conceptos, elaborados por humanos, tienen en cuenta todos los sentidos de la percepción, además de otros conceptos sociales más complejos como 'sirve para cobijarse del frio', etc. que nuestro modesto experimento no

puede abarcar. Quizá el producto del aprendizaje realizado en nuestro experimento sea aún demasiado burdo para ser considerado como concepto, pero podemos afirmar que su contenido ya no es 'no conceptual'.

Concluyendo, en este epígrafe ha tratado de mostrar la distinción entre conceptual / no conceptual desde una perspectiva representacional. Esta visión nos será útil para encuadrar los elementos significativos del experimento (v. supra 2.3, p. 43), que son de carácter representacional, en las distintas teorías de conceptos (v. infra 3.2, p. 76) y en las distintas ontologías (v. infra 3.3, p. 87) que existen para humanos. También será aprovechada en el capítulo siguiente (v. infra 4.1, p. 104) para elaborar una alternativa representacional del experimento descrito en esta tesis. El carácter conceptual desde una perspectiva disposicional será tratado más adelante. Comenzaremos con la visión ontológica de los conceptos como habilidades (v. supra 3.3, p. 93) y seguiremos en el capítulo siguiente presentando una alternativa disposicional a los posibles conceptos derivados del experimento (v. supra 4.2, p. 121).

En resumen, admitimos que en nuestro experimento hay elementos "análogos" a los contenidos no conceptuales de los estados de percepción. Pero también otros estados, a los que Davidson y para humanos denomina estados de creencias, y que en nuestro experimento definiremos como aquellos que soportan lo aprendido por la máquina, que son de contenido 'conceptual'. Estos estados, los que soportan lo aprendido en la percepción, según Roskies no pueden ser del mismo contenido que los estados de percepción, 'no conceptuales'. En este sentido, el contenido no conceptual consistiría en: la imagen, la matriz numérica que representa a la imagen en el espacio RGB o HSV, el vector de descriptores de la imagen. Y el contenido conceptual estaría compuesto por el vector beta y el límite de la equiprobabilidad de cada concepto, junto con su estructura de características virtuales y la función de predicción. Excluyo deliberadamente la red de inferencias y cualquier posible sistema simbólico que se pueda desarrollar a partir de los resultados del experimento por estar fuera del alcance de esta tesis.

3.2 *Las principales teorías de conceptos*

Una cuestión preliminar es por qué son relevantes para esta tesis las teorías empíricas o filosóficas sobre la estructura de los conceptos humanos, si de lo que estamos hablando en realidad es de un *programa de ordenador*. Sencillamente porque el objetivo de la máquina no es aprender a clasificar instancias sin más, sino aprender tomando como referencia la clasificación realizada por un humano, y en este sentido, hay que conocer lo que aquellas teorías sostienen acerca de los conceptos para poder dar respuesta a cómo hacer aprender a una máquina como lo haría un ser humano, si es que eso es posible. Por eso los patrones humanos son significativos aquí, independientemente de que su reelaboración posterior quede enmarcada en el campo de la inteligencia artificial.

Dicho esto, conviene aclarar, primero, cuándo estamos ante un concepto 'simple'. Por 'simple' me refiero a que se expresa con una sola palabra del léxico de un idioma ('soltero', 'gato',...) y que suelen llamarse conceptos *léxicos*. Inmediatamente surge un problema cuando queremos expresar un concepto con más de una palabra, por ejemplo 'gato_negro'. ¿Estamos ante un concepto estructurado en dos conceptos simples o ante un solo concepto con una característica específica? ¿Podemos hablar de conceptos atómicos y conceptos con estructura? Y si admitimos los conceptos estructurados, tendremos que analizar su estructura, y hablaremos de "modelos de contención", cuando el concepto estructurado contiene los conceptos simples, es decir, una ocurrencia del concepto estructurado implicaría instancias diferentes en sus conceptos simples. Por otro lado, hablaremos de "modelos de inferencia", por ejemplo, de rojo inferimos color, y frente al concepto complejo 'mesa_roja' podemos inferir 'mesa_de_color_rojo' (Margolis & Laurence 1999).

Aunque la mayoría de las discusiones sobre conceptos tratan de los conceptos léxicos, en la presente tesis solamente podremos hablar de conceptos léxicos en el sentido de que los conceptos a clasificar son una yuxtaposición de palabras con un sentido único. Aunque esto deja abierta la posibilidad de explorar la estructura del concepto y su descomposición en conceptos más simples, o la catalogación en familias (por ejemplo, en el concepto 'timeofday_day', parece lógico hablar de la familia 'timeofday', y dentro de ella el concepto 'day'), este no será el enfoque por el que tomaremos partido. Y no lo haremos porque, como se discutió en el experi-

mento sobre la elección de imágenes de entrenamiento, no se apreciaron diferencias significativas entre la elección de imágenes negativas dentro de la familia (baseline 3) o fuera de la familia (baseline 2) a la que pertenece el concepto (v. supra p. 42).

En cuanto a las teorías sobre la estructura de los conceptos, Margolis y Laurence (Margolis & Laurence 1999) distinguen, entre otras, la teoría clásica, la teoría de los prototipos y la teoría "Teoría". El punto de partida, en todo caso, debe ser la teoría clásica, ya que según estos autores, de una u otra forma, todas las demás reaccionan ante ella, sea para ampliarla o para rebatirla.

(a) La teoría clásica.

Esta teoría sostiene que la mayoría de los conceptos, especialmente los léxicos, tienen definiciones de estructura, lo que significa que la mayoría de los conceptos codifican las condiciones necesarias y suficientes para su aplicación. Por ejemplo, el concepto 'soltero' podemos pensarlo como un complejo mental que especifica las condiciones necesarias y suficientes para que algo sea soltero. En este caso, podría tener representaciones de 'no_casado', 'masculino', 'adulto'. Cada uno de estos componentes incluye una condición que algo tiene que cumplir para ser incluido en 'soltero'; y cualquier cosa que las cumpla todas, será un soltero. Los componentes, que llamaremos características, apuntan hacia una interpretación semántica composicional.

En suma, la teoría clásica sostiene que los conceptos son representaciones complejas integradas por representaciones estructuralmente más simples. Y si estos componentes más simples son también representaciones complejas, se pueden seguir descomponiendo de igual modo hasta encontrar representaciones primitivas. Estas representaciones primitivas suelen remitir a cualidades sensoriales o perceptivas simples, cuando hablamos de conceptos cotidianos como los que se utilizan en nuestro experimento, aunque hoy en día ni siquiera los empiristas más recalcitrantes defenderían la tesis de que cualquier concepto es definible en un vocabulario puramente sensorial. Piénsese en conceptos como 'electrón' o 'tiempo'.

En todo caso esta teoría ha sido la predominante a lo largo de la historia del pensamiento y ha comenzado a ser desafiada a partir de 1950 aproximadamente en la filosofía y de 1970 en la psicología. Su atractivo es la sencillez y la potencia explicativa-descriptiva. Sencillez porque basta comprobar que se cumplen todas las características para clasificar una instancia dentro del concepto. En este sentido,

clasificar un ítem bajo un concepto tendría una justificación epistémica por el hecho de que sabemos que este elemento tiene las características que definen el concepto, y esta sería su potencia explicativa-descriptiva. Otro atractivo de la teoría es su capacidad para realizar inferencias analíticas, por ejemplo, si S es soltero, puedo inferir que S es hombre, porque hombre es una característica de soltero. Además las inferencias que pueden realizarse con los conceptos pueden dar lugar a categorías o clases.

En relación a nuestro experimento, la teoría clásica podría dar cobertura a lo que hemos llamado características virtuales, es decir, la agrupación de descriptores sobre los que realizaremos la regresión, y que proporcionarán tras esta, la probabilidad de que, según la característica, una instancia contenga un determinado concepto. Pero hay que tener en cuenta que las características que utilizaremos son vectores numéricos que no tienen un significado por sí mismos, como ocurre con 'soltero', 'no_casado' y 'hombre', y por tanto, es una diferencia fundamental respecto de las características planteadas en la teoría clásica. Por el hecho de carecer de significado, las características virtuales pierden el principal atractivo de la teoría clásica, que es su poder explicativo-descriptivo.

(b) La teoría "del prototipo".

En la década de los setenta del siglo pasado surge la primera alternativa seria a la teoría clásica, la "teoría del prototipo", como un planteamiento para dar solución a ciertos problemas que la psicología empírica había planteado. Dicha alternativa es una versión idealizada de una amplia clase de teorías que en síntesis sostienen que la mayoría de los conceptos, incluidos los léxicos, son representaciones complejas cuya estructura codifica un análisis estadístico de las propiedades que poseen sus miembros. De esta forma, donde la teoría clásica necesitaba satisfacer una serie de condiciones necesarias y suficientes para que una instancia satisfaga un concepto, la teoría de prototipos admitirá que también ejemplifica un concepto una instancia que satisfaga una serie más reducida de características teniendo en cuenta, además, que algunas pueden tener diferente peso.

Esta teoría psicológica encaja con la sugerencia de Wittgenstein de que las cosas que instancian un concepto pueden ser muy diversas, y que a veces comparten semejanzas de detalle, pero otras veces se trata de similitudes generales, más bien un "aire de familia" (Wittgenstein 1953[1958] [capítulo 6], pág 32). En cuanto a

las investigaciones empíricas, son especialmente reseñables los trabajos de Eleanor Rosch realizados a partir de los años setenta del pasado siglo (v. Vega de 1984, cap. 7).

Un concepto que ilustra bien la diferencia con la teoría clásica es el concepto "juego", propuesto como ejemplo por Wittgenstein. No podemos dar una definición basándonos en sus características ("número de jugadores", etc.) porque habrá instancias que cumplirán unas y no otras, de modo que no habrá un núcleo de características que cumplan todas aquellas cosas que consideramos "juegos". Luego podemos concluir que el concepto juego no es definible según la teoría clásica. En la teoría de prototipos, que afirma que los conceptos no tienen definición de estructura, sino que sus propiedades se superponen de manera que muestran un espacio de similitud, sí que podemos encontrar una función que permita decidir si una determinada instancia puede incluirse bajo el concepto juego. Esto ayuda a sortear este problema de la falta de definición "cerrada" de algunos conceptos y el de la analiticidad, a saber, si el conjunto de condiciones necesarias y suficientes deja cerrado de un modo concluyente lo que instancia el concepto y lo que el concepto es, entonces un cambio en una sola de las condiciones genera un nuevo concepto y modifica su extensión drásticamente. La teoría del prototipo encaja mejor, pues, con la idea de que el concepto puede ser revisable en un futuro.

Por otro lado, tanto la teoría clásica como la de prototipos construyen un concepto mediante el ensamblaje de sus características y, cuando estamos ante conceptos de la vida cotidiana, en ambos casos se asume que las características están conectadas con propiedades perceptibles. La principal diferencia es que en la teoría de prototipos las características de un concepto expresan propiedades estadísticamente importantes, por esto no es necesario que una instancia cumpla todas las características. De esta forma se ofrece un tratamiento para la categorización mucho más flexible y robusto basado en funciones de similitud[15].

[15]En el modelo propuesto por Medin y Schaffer (Medin & Schaffer 1978) "*Context Model*" ; primero se determinan las características que pertenecen al núcleo, es decir, aquellas que siempre han de estar presentes en una instancia para poder pertenecer al concepto (en este sentido, el núcleo se trataría como en la teoría clásica); después se determinan aquellas características secundarias que influyen en la determinación de la pertenencia al grupo. Para establecer el grado de similitud de una característica de una instancia respecto al modelo ejemplar o prototipo, si pertenece al núcleo se le asigna un uno, si no, se calcula cuantas instancias de todo el universo considerado presentan el mismo valor respecto al total de instancias, y se le asigna el valor de este cociente. La medida de

No obstante, un problema al que se enfrenta esta teoría es el problema de la tipicidad. ¿Por qué para el concepto 'ave' es más típica una instancia del concepto 'paloma' que una del de 'avestruz'? Para los humanos puede que sea algo relacionado con la velocidad de procesamiento, es decir, rápidamente procesamos el núcleo de características de 'paloma' para identificarla con 'ave', en cambio, nos cuesta más ese procesamiento para 'avestruz'. Aunque así no se evita que la cuestión resurja en los siguientes términos: ¿por qué procesamos más rápidamente una imagen de paloma que una de un avestruz a la hora de adscribirlas a la clase de las aves? Esto plantea el problema de "si los juicios de tipicidad hechos sobre una instancia para clasificarla bajo un concepto no son el grado de pertenencia , entonces

similitud total es el producto de todas las características. Por ejemplo, consideremos que tenemos cuatro características que pertenecen al núcleo del concepto y una que no pertenece. Supongamos que para la instancia a clasificar, la característica que no pertenece al núcleo presenta una proporción de $\frac{1}{2}$, es decir, de todas las instancias que pertenecen al concepto, la mitad presentan esta característica. Entonces, la similitud de esta instancia respecto al modelo sería $1 * 1 * 1 * 1 * 0,5 = 0,5$. Este modelo admite mayor flexibilidad que el modelo clásico en el sentido de que, para que una instancia pertenezca al concepto, en el modelo clásico, aquellas características que no pertenecen al núcleo, deben tener un valor de uno, si no, sería rechazada. Mientras en este modelo, se establece un valor de similitud con el modelo ejemplar o prototipo que permite admitirla en el concepto con un valor menor.

Posteriormente Nosofsky y Palmeri (Nosofsky 1992 y Nosofsky & Palmeri 1997) ampliaron este modelo hacia el "*Generalized Context Model*". En este, la medida de similitud es más compleja y considera más factores. Por una parte calcula la distancia euclídea para cada característica entre la instancia a clasificar y el prototipo. Después se le asigna un peso (w) según la relevancia de la característica respecto al concepto. Y por último, se clasifica en el concepto que mayor similitud tenga.

$$d_{ij} = \sqrt{\sum_m w_m |x_{im} - x_{jm}|^2}$$

Hay que tener en cuenta que una mayor distancia supone una menor similitud. Para transformar una distancia en una similitud hay muchos procedimientos, desde calcular la inversa de la distancia:

$$s_{ij} = \frac{1}{d_{ij}}$$

la forma exponencial decreciente propuesta por Shepard (Shepard 1987):

$$s_{ij} = e^{-c*d_{ij}})$$

etc.

En este sentido, la teoría de prototipos es capaz de clasificar correctamente instancias atípicas al asignar pesos ponderados a sus características, que serían rechazadas por la teoría clásica. Es, en definitiva, más robusto porque no necesitamos nuevos conceptos para clasificar instancias atípicas.

¿qué son?". La respuesta no parece simple.

Tal como se vio en la nota 15, p. 79, la similitud es una magnitud que nos dice lo semejante que es nuestra instancia respecto al prototipo, con un rango de valores [0, 1], siendo 0 ninguna similitud y 1 el máximo de similitud. En este sentido, la instancia-modelo o prototípica es el más semejante al prototipo, y por esa cuestión la similitud vale 1. Además, si consideramos al prototipo como el concepto mismo a efectos métricos, y según Smith y Medin (Smith & Medin 1981) la tipicidad de un ejemplar es una medida de la semejanza entre dicho ejemplar y su prototipo, entonces es razonable pensar que un juicio de tipicidad sobre una instancia sea un juicio sobre la semejanza que esta tiene con el prototipo, que a efectos métricos corresponde al concepto. ¿Afirmaríamos que los juicios de tipicidad son el grado de pertenencia al concepto? En todo caso, el prototipo codifica una serie de propiedades del concepto, pero no es el concepto. Podría darse el caso de juicios de tipicidad donde una instancia, que no pertenece al concepto, presente un grado de pertenencia a un concepto mayor que otra, que si pertenece. Puede darse si la medida de similitud ha sobreestimado el peso de una característica respecto a otras, o, que las características fundamentales no estén representadas en el prototipo. Por ejemplo, si tenemos un prototipo del concepto "naranja" solamente con las características color:naranja y forma:esférica, puede ocurrir que una pelota naranja presente un grado de tipicidad (o similitud) mayor que una naranja deforme verde. En este caso, no podríamos afirmar que el juicio de tipicidad sea el grado de pertenencia de la instancia al concepto. En este sentido, para poder afirmar que los juicios de tipicidad son el grado de pertenencia al concepto, debemos exigir que el prototipo tenga las características necesarias y suficientes para representar al concepto, pero además que estas estén bien ponderadas. Las características necesarias conformarían el núcleo y, junto con las restantes, acomodarían a las instancias atípicas. En todo caso, la teoría de prototipos sigue reduciendo el concepto a una representación composicional de características.

Este problema ha llevado a sugerir teorías duales de conceptos (Smith & Medin 1981) que pretenden descomponer el problema de la clasificación en, primero, un procedimiento de identificación rápida, que categoriza la instancia, y segundo, un proceso más reflexivo donde se encuadra mejor la instancia.

De forma lógica, en el primer proceso, el humano debería computar el núcleo

de características, para descartar rápidamente las instancias que no lo cumplen. En un segundo proceso más reflexivo computaría las restantes, además por orden de importancia (el peso de la característica), incrementando así el valor de su tipicidad hasta alcanzar el umbral necesario para considerarse como instancia del concepto, o ser rechazada tras agotar el cómputo de todas las características.

Pero el humano tiene otra lógica para secuenciar el cómputo. Es decir, para un concepto determinado, le puede ser más fácil computar características con menos peso que las características necesarias, y por este motivo las computa antes. Por ejemplo, para el concepto 'ave', la instancia 'paloma' el humano la identifica más rápidamente que la instancia 'avestruz'. Este caso puede ser explicado en esta teoría dual porque la característica 'volar', que no pertenecería al núcleo del concepto, es computada antes que otras que sí pertenecen al núcleo.

Por tanto, la primera fase de identificación rápida apela al cómputo de las características que para el humano son más comunes y rápidas, y en la segunda fase más reflexiva, se computarían las características menos comunes y más lentas para el humano. No se trata de una secuenciación lógica respecto al concepto, sino respecto a la facilidad de cálculo en relación a las capacidades cognitivas humanas.

En nuestro experimento todas las características se computan al mismo tiempo y a la misma velocidad. Además el cálculo de tipicidad se completa antes de someterlo al umbral (límite de equiprobabilidad) que determina la pertenencia al concepto. El caso es distinto en humanos, y para los humanos clasificadores de la base de datos del experimento, sometidos a presión temporal, puede que les afecte acortando la duración de la segunda fase y cometiendo errores de clasificación.

Por desgracia, esto no resuelve el problema de la ignorancia y error. Los mismos argumentos que Kripke y Putnam (Kripke 1972; Putnam 1975) presentan en este sentido para evaluar la teoría clásica, no se resuelven en la teoría de prototipos. Hay errores debidos tanto a la falta de características de un concepto como a características que se muestran erróneas por nuevos avances en la investigación de un concepto determinado. Así, estos autores afirman que se puede tener un concepto sin representar las condiciones necesarias y suficientes para su aplicación y esta carencia es causa de errores en la tarea de clasificación. Aún suponiendo que para un concepto se puedan encontrar las características necesarias y suficientes que lo representen, nos podemos encontrar con características mal definidas,

o características poco conocidas que mantengan el problema de la ignorancia y el error. Este es un problema estructural que persiste debido, entre otros, al intento de reducir un concepto a una composición de sus características. Y en la teoría de prototipos otra fuente de errores es, además de las anteriores, la ponderación de las características que lo componen.

De hecho, en los agentes humanos las instancias atípicas de un concepto pueden desembocar en errores de clasificación. En este sentido, los errores de clasificación de los sujetos humanos en el experimento, que son quienes realizan la clasificación de partida y a quienes se supone veracidad, pueden ser parte de la explicación de las tasas de errores que comete la máquina. La máquina no haría más que reproducir los errores cometidos previamente por el humano. Para poder probar estas afirmaciones necesitaríamos un juez externo que determinara cuáles son los errores cometidos por los humanos clasificadores, y si estos, realmente se reproducen en la máquina. Si así fuera, la máquina sería un simulador de la actividad humana de clasificación. Entonces, ¿qué juez?, ¿otro humano?, ¿acaso este estaría libre de los errores comentados en los párrafos anteriores? Entendemos que no y por este motivo no se ha elaborado una traza de seguimiento. Tampoco nos interesa probar que la máquina es un simulador de clasificación humana de imágenes. Simplemente se ha aceptado como verdadera la clasificación realizada por los agentes humanos para el aprendizaje y la clasificación que realiza la máquina.

En nuestro experimento, las características están ponderadas como promedio, para todos los conceptos. Además, su valoración es de escala numérica y han sido extraídas empíricamente de la imagen a la que representan. Aunque este proceso se libre del problema de la ignorancia y el error, sí podríamos encontrarlo en la determinación de las características del núcleo y en la ponderación de las demás características referidas a cada concepto en particular.

También podríamos decir que la teoría prototípica se adecua mejor al experimento porque, como vimos en la parte práctica, este trata los descriptores de las características virtuales no como una condición necesaria y suficiente para que una imagen instancie un concepto, sino que las trata estadísticamente mediante medidas de similitud –en nuestro experimento, la aplicación de la función de distribución derivada de la regresión logística-, que nos ofrecerá una medida para cada característica virtual que conecta la imagen y el concepto. No obstante, cabe destacar

que el sentido de similitud en la máquina y en el humano son diferentes. Mientras para la máquina la similitud esta basada sobre distancias entre características estrictamente físicas de la imagen, para los humanos las distancias lo son entre características que tienen un significado perfectamente delimitado para el humano. Es decir, para las características utilizadas por los humanos, estos deben interpretar la imagen, y para ello puede que no sea suficiente con la información estrictamente física de la imagen.

Por otra parte, la forma de ponderar la similitud general será mediante la media aritmética de las probabilidades de las características virtuales. Esta medida se ha tomado por la experiencia del autor de esta tesis en los experimentos realizados durante los años de investigación dentro de este campo en inteligencia artificial (Benavent et al. 2010, y 2012; Benavent et al. 2013; Granados et al. 2011; Castellanos et al. 2011, y 2012; de Ves et al. 2016). Una posible ampliación en este sentido, sería encontrar qué características virtuales son más relevantes para un concepto determinado, y poder computarlas ponderadamente, aunque en la presente tesis no abordaremos este punto. En la parte de las conclusiones retornaremos sobre el problema de la tipicidad.

Por último, como ya discutimos anteriormente, es sugerente considerar al centroide de las imágenes positivas como un 'prototipo virtual', es decir, sin que corresponda a ninguna imagen. (v. supra 2.3.4, p. 47 y 2.4, p. 56) , el centroide podría ser satisfecho por más de una imagen que, además, no tuvieran ningún contenido semántico para un sujeto humano. Por eso la identificación del centroide con un 'prototipo virtual' hay que tomarla con mucha cautela. Además de los problemas sobre el significado de las características virtuales discutidos, hay otras consideraciones. En primer lugar hay un centroide de imágenes positivas y otro de negativas. Su función es ayudar a encontrar las imágenes más cercanas a ellos mismos (con métrica euclídea sobre sus descriptores) para elegir las imágenes positivas y negativas que servirán de base de entrenamiento al clasificador automático y ahí acaba la función de los centroides. Segundo, para la teoría de prototipos, la similitud, como hemos visto, está relacionada con las distancias que separan a cada característica de la instancia con su homologa del prototipo. En nuestro experimento, esta similitud, se calcula en términos de una función que no ha tenido en cuenta los centroides, sino las imágenes más próximas a estos. Además, si calculamos la similitud del

centroide al concepto, en general no sera uno[16]. Con todos estos matices, el centroide de positivas es aún el mejor candidato a considerarse como un 'prototipo virtual' del concepto.

Entonces, si la teoría de prototipos necesita los parámetros de las características de este para poder establecer una medida de similitud con la instancia que se pretende clasificar, ¿dónde se encuentran estos parámetros?, ¿son estos parámetros realmente el prototipo? Aunque son de una naturaleza distinta, podemos identificar estos parámetros con el vector beta, ya que son la fuente de datos que nos permitirá calcular la similitud de la instancia con el concepto[17]. Además, como ya se apuntó en la sección anterior, el contenido del vector beta representa lo aprendido en el proceso de clasificación de un concepto determinado. Pues bien, ¿sería el vector beta un posible candidato a prototipo del concepto? Intentaremos responder esta cuestión en capítulos posteriores.

(c) Otras alternativas.

¿Qué decir del resto de opciones (teoría-Teoría, neoclásica, atomismo conceptual) respecto a la estructura de los conceptos? Solamente discutiremos brevemente la teoría-Teoría porque la neoclásica y el atomismo conceptual aportan poco contenido adicional al ya discutido para nuestro experimento.

Susan Carey –una de las principales defensoras de la "teoría-Teoría", junto con otros autores como Gregory Murphy, Douglas Medin o Alison Gopnik (v. Margolis & Laurence 1999, 43 y ss.) propone que los conceptos sean identificados por los roles que desempeñan en las teorías. La idea es que algunos cuerpos de conocimiento tienen unas características que los identifican con las teorías científicas, y los conceptos que se representan en estos cuerpos de conocimiento se individualizan por sus roles cognitivos en sus respectivas teorías mentales. Esta sería la llamada teoría "Teoría" de los conceptos.

[16]Como ya hemos mencionado, el prototipo es el más similar al prototipo con lo que se espera que la medida de similitud sea uno. En general no se cumple porque la regresión intenta ajustar a uno las imágenes de entrenamiento, no el centroide.

[17]Los parámetros del vector $\overrightarrow{\beta}$ son de una dimensión mayor que el espacio de descriptores. Como el primer parámetro es solamente de ajuste, podemos considerar el vector $\overrightarrow{\beta^*}$ donde se ha eliminado el parámetro β_0. Si consideramos a este vector como centroide, en la función que calcula la similitud de este con el concepto, $\overrightarrow{X} = \overrightarrow{\beta^*}$, y deberia cumplirse que

$1 = \frac{1}{1+e^{-(\beta_0+\overrightarrow{\beta_i}*\overrightarrow{X_i})}} = \frac{1}{1+e^{-(\beta_0+\overrightarrow{\beta_i}*\overrightarrow{\beta_i^*})}}$

que en general no se cumple.

3. DISCUSIÓN DEL EXPERIMENTO EN RELACIÓN AL DEBATE CONTEMPORÁNEO SOBRE LA NATURALEZA DE LOS CONCEPTOS

Los conceptos, según esta teoría, se extraerían de un cuerpo de conocimientos implícito en las imágenes de una base de datos. Supongamos que disponemos de dos agentes humanos para clasificar las imágenes, un cardiólogo y un adolescente enamorado. Les proponemos que elijan las imágenes que más se asemejan al concepto léxico "corazón". Es muy probable que las imágenes que elija cada uno sean bastante dispares, aunque la elección la hayan hecho sobre la misma base de datos. ¿Quiere decir esto que el concepto "corazón" no es compartido por ambos agentes? Y si no es compartido un concepto, ¿de qué sirve? La explicación que proponen los seguidores de la teoría-Teoría consistiría en que cada agente extrae el concepto "corazón" de un dominio diferente. Este dominio personal del agente, denominado 'teoría', es el responsable del sentido dado al concepto.

Desde el planteamiento de nuestro experimento, esta teoría sobre la estructura de los conceptos es inabordable. Aunque en nuestro experimento podríamos considerar que una determinada base de datos esté directamente relacionada con un corpus de conocimiento, casos extraídos de bases de datos médicas, etc., en general esta no va a ser la posición de partida. Tampoco parece adecuado entender cada base de datos como un corpus de conocimiento '*per se*' sobre el que podamos extraer conceptos atendiendo al papel que estos desempeñan, ya que en general un conjunto de imágenes sin más no tiene por qué poseer una estructura relacional que nos haga pensar en ella como un dominio de conocimiento, o al menos, como una base de entrenamiento adecuada para un concepto[18]. Además, en principio, lo que nos va a interesar es trabajar con las características virtuales.

Se puede objetar que los conceptos presentados en el experimento sí son dependientes de la base de datos utilizada, y en este sentido, los conceptos han sido extraídos de un marco teórico específico, entendiendo este como el dominio de la base de datos utilizada. Esta objeción explicaría por qué una vez obtenido el modelo del concepto puede fallar al aplicarlo sobre una base de datos distinta. Simplemente el concepto no estaría ubicado en su contexto teórico propio. Pero esta identificación directa entre base de datos y contexto teórico resulta tosca y empobrecedora. No es eso en realidad en lo que están pensando los defensores de la teoría-Teoria. Se apuesta más bien de una caracterización *funcional* del concepto. Por estas razo-

[18]Como vimos en 2.2.3 p. 42, ni siquiera la forma de elegir las imágenes negativas, dentro o fuera de la familia de un concepto, repercutía sustancialmente en los resultados.

nes, pensamos que dicha teoría no sería adecuada para interpretar los resultados de la clasificación hecha por la máquina.

En resumen, al intentar enmarcar el experimento dentro de una teoría sobre la estructura de los conceptos, hemos visto que la que mejor se adapta al contexto y los requerimientos prácticos de nuestro experimento es la teoría del prototipo.

3.3 *El estatus ontológico de los conceptos*

¿Cómo se representa un concepto en nuestro experimento? Se podría admitir que los parámetros de la función que usamos para calcular la probabilidad de pertenencia de una instancia a un concepto determinado son, en realidad, la representación de este concepto. En este sentido, el vector beta junto con el límite de la equiprobabilidad jugarían un papel fundamental para determinar la ontología del concepto que se desprende del experimento. Por otro lado, podríamos admitir también que el conjunto de instancias positivas y negativas que se suministran a la función de regresión para calcular el vector beta son las que contienen implícitamente el significado del concepto, y en este sentido, jugarían también un papel importante en la determinación del estatuto ontológico de los conceptos. Y aún más, una tercera entidad que subyace en la elección automática de las instancias positivas y negativas que se utilizarán para la regresión, es decir, el centroide de positivas y el centroide de negativas para cada concepto, parece que tenga también algo que decir sobre la ontología de conceptos. Relacionaremos estos términos con el marco de la discusión sobre el estatuto ontológico de los conceptos, discusión que plantea tres opciones básicas: los conceptos entendidos como representaciones mentales, como entidades abstractas (los sentidos fregeanos, por ejemplo), o como habilidades.

(a) Los conceptos como representaciones mentales.

Según Margolis y Laurence (Margolis & Laurence 2014) para quienes sostienen que los conceptos son representaciones mentales, el pensamiento se produce en un sistema interno de representación, donde los conceptos son entidades psicológicas. Admiten que tanto las creencias como los deseos y otras actitudes proposicionales son símbolos internos para los procesos mentales en los humanos. Además, dotan a sus símbolos de la característica causal-funcional típica de las creencias. Por ejemplo, aplicando funciones entre símbolos, si S cree que A es mayor que B y también

que B es mayor que C, estas dos creencias **causan** la creencia en S que A es mayor que C. Las representaciones que figuran en las creencias de S estarían compuestas por representaciones más básicas, y en este sentido, los conceptos serían las representaciones más básicas de las creencias de forma que (A, B, C, mayor que) serían símbolos, y las creencias de S representan lo que los símbolos hacen en virtud de, primero los contenidos de los símbolos, y segundo, de cómo se organizan.

Podemos considerar a John Locke o a David Hume como defensores iniciales de esta posición representacionalista. Ellos hablaron de unas representaciones básicas, unas ideas simples, que catalogaron como imágenes mentales. Actualmente se sostiene que el sistema interno de representación no tiene por qué estar compuesto exclusivamente por imágenes mentales, ya que gran parte del pensamiento humano no se basa en ellas. Dicho sistema es similar al lenguaje –posee una sintaxis y una semántica composicional- de ahí que se haya hablado del "lenguaje del pensamiento". Uno de los más conocidos representantes de esta posición sería Jerry Fodor.

En contra de la identificación entre conceptos y representaciones mentales, puede objetarse que puede haber diferentes representaciones mentales de un mismo concepto, es decir, puede que diferentes humanos usen símbolos diferentes para representar un mismo concepto, y aún más, puede que un mismo símbolo sea usado para conceptos distintos. Además aun suponiendo que hubiera acuerdo en los símbolos usados para cada concepto, no queda claro que todos los humanos capten la misma semántica del concepto. Y si, en definitiva, lo que se construye es un tipo de lenguaje al que se pueden aplicar unas reglas de composición, quedaría por determinar aún el contenido de los conceptos.

En esta misma línea Vega (Vega de 1984 , caps. 5, 6, 7) considera que las representaciones mentales podemos verlas como imágenes mentales o como representaciones proposicionales. En el primer caso se podría objetar que no todos los conceptos tienen una imagen que los represente, por lo que necesitará un sistema de codificación diferente -quizá el sistema verbal – e incluso, usar una representación dual como la suma de imagen mental y sistema verbal. En el segundo caso, las representaciones proposicionales pueden ser útiles para representar el conocimiento, mapas semánticos (TCL[19] de Quillian (Quillian 1968)), en concreto los casos

[19] *Teachable Language Comprehender*. Es la forma de organizar el conocimiento conceptual en la memoria de un hablante, que él lo llamó memoria semántica. Consiste en nodos conceptua-

en que un sistema proposicional represente a un concepto, pero todavía quedaría por explicar los conceptos implícitos en las proposiciones. Si volvemos a utilizar un sistema proposicional para explicar cada concepto implicito estariamos ante una regresión que nunca terminaria.

En definitiva, si el concepto es una representación mental y estas -o algunas de estas al menos- se entienden como representaciones proposicionales, entonces habremos avanzado poco en el esclarecimiento de lo que son los conceptos, puesto que las representaciones proposicionales, ¿cómo pueden ser tal clase de representaciones si no contienen -explícita o implícitamente- conceptos o algo parecido?

Una alternativa actual al "lenguaje del pensamiento" viene, por un lado, de los avances en la modelización computacional conexionista. El conexionismo presenta los fenómenos de la mente como procesos que emergen de redes de elementos sencillos interconectados. Y por otro lado, de la teoría de sistemas dinámicos[20]. En estas teorías, la mente computacional fue concebida inicialmente como una mente simbólica al estilo de los ordenadores digitales, es decir, como un procesador serie de información en forma de símbolos discretos, estructurados sintácticamente y manipulados de acuerdo con reglas ajenas al contenido representado -un programa en un lenguaje de programación. Para el cognitivismo clásico [21] , la actividad representacional de la mente estaría en estos símbolos y su manipulación de acuerdo con las reglas formales establecidas. Con los modelos de "procesamiento distribuido en paralelo" (PDP models)[22] se reformuló la noción cognitivista de representación que

---

les conectados por arcos con las propiedades y otros conceptos. Cualquier proposición puede ser representada en este gráfico. Tuvo posteriores desarrollos que no modificaron sustancialmente su estructura (Collins & Quillian 1972) .

[20]Un sistema dinámico en este sentido es aquel que predice un estado futuro como una función lineal o no del estado presente más las entradas N. Es decir, $E_{k+1} = f(E_k) + N_k$. Un autor relevante en la modelización de sistemas dinámicos para física e ingeniería es Katsuhiko Ogata, (Ogata 1987, cap 1)

[21]El cognitivismo clásico comenzó en 1956 con la noción de que todos los sistemas procesadores de información, incluido el cerebro humano, comparten los mismos principios. A partir de la analogía entre la computadora y el cerebro, se consideró apropiado estudiar la mente como si se tratara de un software.

[22]El PDP es un enfoque de red neuronal que destaca el carácter paralelo de procesamiento de las neuronas o elementos de la red. A partir de él se puede generar un marco matemático para trabajar con varios ítems, que podrían resumirse según Rumelhart, Hinton y McClelland (Rumelhart et al. 1986, cap 2): (i) Unidades de procesamiento, representadas por números enteros; (ii) Cada unidad se activa según un vector de funciones dependientes del tiempo. (iii) Una función de salida para cada unidad. (iv)Patrón de conectividad entre unidades (v) Regla de propagación, que es una función

viraría de la discreción de los símbolos a la dispersión en redes conexionistas: en estas últimas las representaciones se hallan no en elementos, sino entre elementos; se trata de representaciones distribuidas que dependen de propiedades sistémicas dentro de redes neuronales artificiales. Si para el cognitivismo clásico el proceso paradigmático era la inferencia reglada, para el conexionismo lo son el reconocimiento de patrones y un tipo de aprendizaje que prescinde de reglas previas: las reglas surgen de y con la experiencia.

En nuestro experimento, sin embargo, pensamos que sí cabe hablar de *una representación mental* del concepto. Por un lado podemos considerar el centroide como una representación del prototipo del concepto, con todos los matices ya expuestos (v. supra 3.2, p. 78). Por otro lado, aunque el vector beta puede verse como unas constantes que intervendrán en una función, conjuntamente con las características virtuales de la imagen que se ha de clasificar, para determinar la probabilidad de que esta tenga el concepto léxico, es decir, parte de una función; también, de alguna manera, *este vector beta puede considerarse como una representación abstracta del concepto*, y en este sentido, el contenido de beta, y/o la salida de la función de clasificación podrían utilizarse para arrancar procesos de "pensamiento" en la máquina.

Daniel Dennett ha sido uno de los defensores más conocidos de la idea de que podemos atribuir legítimamente actitudes proposicionales a un ordenador aunque carezca de las representaciones a las que apelaríamos para explicar o predecir la conducta de un humano (Dennett 1987) . Así, en el ajedrez decimos que el jugador piensa en enrocar el rey y sacarlo del centro del tablero, y lo mismo podría decirse, según Dennett, del programa de ordenador. Hemos de advertir, no obstante, que aquí nosotros no nos estamos planteando si una máquina piensa o si tiene "intenciones", ni siquiera si ve. Nuestro objetivo es más modesto y tiene que ver con la adquisición de conceptos "visuales".

El núcleo de nuestro experimento, el sistema de aprendizaje, se lleva a cabo mediante regresiones logísticas. La forma de obtener los parámetros beta se podría encuadrar dentro del conexionismo, donde los parámetros se aprenden por la expe-

a la salida de las unidades. (vi) Regla de activación de la unidad, que combina las entradas y la propagación anterior. (vii) Regla de aprendizaje, que modifica los pesos de las entradas basándose en la experiencia. (viii) Entorno para el aprendizaje. Ambiente donde debe operar.

riencia (de las imágenes de entrenamiento), sin patrones previos, y en este sentido, los vectores beta junto con el límite de la equiprobabilidad serían la habilidad de clasificar aprendida para el concepto tratado. El vector beta y el límite de equiprobabilidad serían la representación de la habilidad de clasificación para cada característica virtual de cada concepto. En nuestro experimento, las neuronas[23], ya entrenadas, integrarían los parámetros de su vector beta junto con los descriptores de imagen en la entrada, después aplicaría la función logística y se obtendría el valor de similitud de la característica virtual de la imagen respecto a la misma característica del concepto. Lo aprendido del concepto, que involucra una habilidad en la clasificación, son los vectores beta y el límite de equiprobabilidad.

Por otra parte, no tiene sentido proponer las imágenes positivas y negativas como representaciones mentales del concepto, ya que ninguna de ellas es el producto final de la elaboración del concepto, sino las bases para su elaboración, en todo caso. En definitiva, admitiendo que los centroides sí son un producto elaborado para la determinación de un concepto, y sí podemos considerarlo como representante del conjunto de imágenes positivas por el hecho de ser su centro euclídeo, sería un firme candidato para ser concepto. Pero en el proceso de clasificación no es determinante, ya que para este sólo se tiene en cuenta el vector beta y el límite de la equiprobabilidad.

(b) Los conceptos como entidades abstractas.

Aunque en psicología es usual suponer que los conceptos son representaciones mentales, a menudo la filosofía ha pensado los conceptos como entidades abstractas. Platón podría considerarse un precedente lejano, pero más modernamente algunos filósofos han entendido los conceptos como sentidos fregeanos. Según sus defensores los conceptos son objetos abstractos, y no objetos o representaciones mentales, que median entre el pensamiento y el lenguaje por un lado, y los referentes por el otro. Una expresión sin referente, como 'Pegaso', todavía tiene un sentido. Análogamente, el mismo referente puede estar asociado con diferentes expresiones (por ejemplo, "Superman" y "Clark Kent") porque son portadores de sentidos diferentes. Los sentidos son más exigentes que los referentes. Cada sentido

[23] En nuestro caso la neurona es la propia función de distribución, vista como un integrador y a su salida una función sigmoide seguida de un mecanismo de ponderación de media aritmética con todas las características virtuales y de un umbral de disparo representado por el límite de equiprobabilidad.

tiene una perspectiva única de su referente, un exclusivo modo de presentación. Las diferencias en el contenido cognitivo se remontan a las diferencias en los modos de presentación. Es por esta razón que el pensamiento de que "Superman es Clark Kent" no es trivial, a diferencia de lo que ocurre con "Superman es Superman" . Los filósofos que adoptan los conceptos como sentidos hacen especial hincapié en esta característica.

Para la presente tesis, por una parte, al ser los conceptos mediadores entre los pensamientos y el lenguaje, los agentes humanos que catalogan la base de datos lo hacen tras proyectar sobre las imágenes de la base de datos el concepto sostenido en la mediación entre la expresión escrita y su propio pensamiento. Por otra parte, esta misma expresión escrita puede relacionarse con diferentes referentes, dependiendo del contexto visual de la imagen. En este sentido, en los humanos, la expresión escrita 'timeofday_day', primero, mediará entre los pensamientos de lo que el individuo entiende por esta expresión escrita, y segundo, proyectará este pensamiento sobre las imágenes a clasificar de tal manera que, tras la interpretación de una imagen de la base de datos, determinará si existe o no este pensamiento, aunque encuentre diferentes referentes.

En concreto, para interpretar nuestro experimento esta forma de concebir los conceptos no nos va a ser útil. Si intentamos obtener los referentes que tiene la máquina sobre la expresión escrita del concepto, nos vamos a encontrar que van a depender del conjunto de imágenes positivas y negativas que le suministremos; en esencia, todas ellas. Nos encontraríamos ante conceptos léxicos con muchos referentes. ¿Cómo extraer el sentido fregeano del conjunto de referentes? ¿Podemos admitir el vector beta como un objeto abstracto que representa al concepto fregeano? Aceptemos que sí. En tal caso, el contenido del concepto dependerá siempre del conjunto de imágenes positivas y negativas que usemos. Solamente con que cambie una imagen, entre millares de ellas, el contenido variará. Y no parece que el sentido fregeano esté expuesto a este tipo de variaciones, que reclaman rasgos como mayor flexibilidad y autocorrección.

Además, siendo el concepto el sentido fregeano, y no el referente, se debería analizar la imagen a clasificar para encontrar tal sentido en ella. Es muy fina esta concepción y, para los objetivos del experimento, estaríamos descargando toda la responsabilidad de la captura del sentido fregeano en el conjunto de imágenes po-

sitivas y negativas elegidas para el entrenamiento. En resumen, el concepto (el que elabora el ordenador, si es que elabora alguno) va a depender de un modo decisivo del conjunto de imágenes de entrenamiento.

Concluyendo, si admitimos este punto de vista ontológico sobre los conceptos, por un lado admitimos en los humanos clasificadores diferentes interpretaciones de una misma imagen, y como consecuencia posibles clasificaciones distintas de la misma imagen bajo conceptos distintos; por otro lado, pueden haber diferentes interpretaciones de un mismo concepto léxico, que dependerán del sentido que le confiera el humano clasificador; y por último, la dependencia del sentido del conjunto particular de imágenes de entrenamiento pone en duda la justificación universal, por así decirlo, de los resultados del experimento, algo que parece exigido o al menos presupuesto por la caracterización de los conceptos como objetos abstractos. En suma, no será este un planteamiento ontológico adecuado para nuestro experimento.

(c) Los conceptos como habilidades.

Quienes sostienen que los conceptos son capacidades, y en este sentido realizar algo exitosamente sería la manifestación visible de que se posee la capacidad/habilidad, plantean una perspectiva bien distinta. Para estos autores ya no cabe afirmar que los conceptos son representaciones mentales particulares, imágenes mentales, o palabras en un "lenguaje del pensamiento"; son habilidades propias de los agentes cognitivos.

La razón más importante para la adopción de este punto de vista es un profundo escepticismo acerca de la existencia y la utilidad de las representaciones mentales, un escepticismo que se remonta a Ludwig Wittgenstein (Wittgenstein 1953[1958]) y que podría rastrearse aún más atrás, en el pragmatismo americano de finales del XIX. Uno de los argumentos más influyentes en este sentido afirma que las representaciones mentales son explicativamente inactivas debido a que reintroducen las misma clase de problemas que supuestamente deben explicar. Por ejemplo, Michael Dummett advierte contra el intento de explicar el conocimiento de una lengua por primera vez en el modelo de conocimiento de un segundo idioma. En el caso de una segunda lengua, es razonable suponer que la comprensión del lenguaje consiste en la traducción de sus palabras y frases, en palabras y frases de la propia lengua materna. Pero de acuerdo con Dummett, no se puede pasar a traducir palabras y

frases de una lengua materna a un lenguaje mental previo. "O, es que realmente no tiene sentido hablar de un concepto. Todo lo que puedo pensar es en alguna imagen, que viene a la mente, que la tomamos como si de alguna manera representara el concepto, y esto nos lleva aún más lejos, ya que todavía tenemos que analizar lo que significa su asociación con ese concepto y en que consiste la imagen" (Dummett 1993, p. 98; citado en Margolis y Laurence (2014), sección 1.2; v. también Laurence y Margolis, 1997). En otras palabras, la representación mental no es, en sí misma, más que otro elemento cuyo significado requiere explicación. De esta manera estaríamos involucrados en una regresión difícilmente aceptable teniendo que recurrir a otro nivel de representación (y así hasta el infinito), o puede ser que también termine con el lenguaje externo y explique su significado directamente, y en tal caso de nada nos ha servido la representación mental.

Nótese que si aceptáramos que los conceptos del experimento tienen en efecto una representación abstracta, digamos el vector beta, estaríamos automáticamente expuestos a la crítica de que la habilidad discriminativa exhibida por la máquina se entiende mejor, se explica, en definitiva, en términos de representaciones subyacentes. Aunque asumieramos, como se ha comentado anteriormente, que, ni las características virtuales ni el vector beta proporcionan información explicativa del concepto en los términos que lo harían las teorías que consideran las características del concepto como explicativas (la teoría clásica y la del prototipo, que vimos en el epígrafe anterior p. 77 y p. 78), sí podríamos explicar esta habilidad en función de los parámetros del vector beta o de las características virtuales. Pero esta explicación carecería de significado semántico.

De todos modos, ***de las tres opciones discutidas en relación a la ontología de los conceptos, esta forma de ver los conceptos "como habilidades" es la que mejor encaja con las coordenadas de nuestro experimento***, ya que lo que se pretende en realidad es instruir a la máquina para que adquiera una habilidad, en concreto, la de clasificar imágenes. Esta opción responde, en principio, a los objetivos planteados por las búsquedas en bases de datos, tanto para el diseño como para la construcción del experimento. No se trata tanto de explicar el concepto, sino de tener la capacidad de clasificar instancias, y en este sentido, podemos ver la adquisición de un concepto como la habilidad para realizar exitosamente este trabajo. Ese es el objetivo del experimento y, en esencia, el vector beta junto con la función de

probabilidad y el límite de la equiprobabilidad constituyen las reglas que subyacen a la habilidad adquirida.

Es importante acotar qué manifestación esperamos, no obstante, por parte del agente que posee el concepto. Podemos distinguir, a este efecto, entre la capacidad *discriminativa* y la capacidad *inferencial*. Por ejemplo, si el concepto 'gato' equivale a la capacidad de discriminar los gatos de los no-gatos, cuando un agente –y aquí es indiferente que sea humano o máquina-, muestre que tiene éxito en dicha tarea en circunstancias complicadas –tal como hace nuestro programa por cierto-, podremos atribuir a dicho agente la posesión del concepto. Visto así los resultados de nuestro experimento confirmarían que el ordenador posee el concepto 'timeof-day_day'. Ahora bien, si la capacidad que se requiere para atribuir la posesión de un concepto es una capacidad inferencial, poco podemos decir a partir de nuestro experimento. Que este vindique una concepción disposicionalista de los conceptos como habilidades dependerá, entonces, de qué actividades se esperan que manifieste quien supuestamente posee el concepto. Si la capacidad discriminativa es suficiente, la respuesta será positiva. Y no solamente cabrá decir que el ordenador posee el concepto, sino también que lo ha aprendido (de un humano, ciertamente). Pero si la capacidad discriminativa no es suficiente (y se exige, por ejemplo, la capacidad inferencial), la respuesta no es ni positiva, ni negativa. Simplemente nuestro experimento no se ha planteado para ver si el ordenador es capaz de realizar tareas inferenciales a partir de lo aprendido (a nivel de discriminación de imágenes, al menos), aunque hay otras ramas en inteligencia artificial que usan como base las inferencias.[24]

[24]Los sistemas basados en conocimiento, por ejemplo, explotan esta característica. Atrás han quedado los sistemas expertos que simplemente capturaban el conocimiento de un experto mediante conceptos y reglas de inferencia, siendo sustituidos por agentes basados en el conocimiento. Básicamente, un agente basado en conocimiento construye una representación del mundo donde debe interactuar y lo guarda en una base de conocimiento. Utiliza también un proceso de inferencia para derivar nuevas representaciones del mundo. Al interactuar, genera nuevo conocimiento comparando el resultado esperado de su proceso de razonamiento con el resultado obtenido de su acción, modificando, si es necesario, su base de conocimiento. Antes de seleccionar cualquier acción, combina el conocimiento general de su base de conocimiento con las percepciones reales para inferir aspectos ocultos del estado del mundo. (Russell & Norvig 2010, cap. 7)

3.4 *Conceptos: ¿innatos o aprendidos?*

Una cuestión que sigue siendo objeto de debate en la actualidad −más en el terreno psicológico que en el filosófico, ciertamente− es si hay conceptos innatos (Griffiths 2009, cap. 5).

Tradicionalmente los empiristas han argumentado que todos los conceptos se derivan de los datos captados por nuestros sentidos, y por ello se construyen a partir de las representaciones sensoriales, de acuerdo con un conjunto de reglas de aprendizaje de uso general. Un ejemplo famoso es el "principio de la copia" de Hume, quien afirmó que los conceptos (las "ideas") se originan a partir de las información sensorial (las "impresiones") y que el contenido de cualquier concepto (idea) debe ser analizable en términos de su base perceptual (impresión). En el terreno filosófico este planteamiento empirista sigue teniendo sus partidarios en la actualidad. Así, Jesse Prinz defiende también una forma modificada de empirismo cuando dice que "todos los conceptos (humanos) son copias o combinaciones de copias de las representaciones perceptuales" (Prinz 2002, p. 108). Lo ampliaremos en el próximo capítulo (v. infra 4.1, p. 104) cuando se discuta sobre el neo-empirismo dentro de la alternativa representacional.

Los innatistas por su parte, entre quienes hay que contar tanto a filósofos (Jerry Fodor) como a la tradición chomskiana en psicolingüística y algunos autores relevantes en psicología cognitiva inicialmente más próximos a planteamientos empiristas (Steven Pinker, por ejemplo), insisten en que la mente humana posee mecanismos innatos de diferenciación en subsistemas específicos de dominios complejos, es decir, existen subsistemas de neuronas específicos que se encargan de dominios complejos como la visión. ¿Se sigue de esto que hay conceptos innatos? Puede que aquellos conceptos cuyas características estén especialmente ligadas a estos subsistemas sean más fáciles de aprender para un humano, pero no parece que deban ser innatos. Por ejemplo, el concepto "triangulo rojo" esta íntimamente ligado a las características físicas de la percepción humana (color y forma), y por tanto parece lógico suponer que será fácil de aprender para un humano. Pero no se sigue necesariamente de ahí que "triangulo rojo" sea un concepto innato en humanos. Ya se discutió el argumento de Roskies en este sentido (v. supra 3.1, p. 74).

Quienes abogan por entender como innato el origen de los conceptos se basan a

menudo en cómo aprende un niño el lenguaje materno, cuando no se le suministran suficientes instancias para formar los conceptos (Laurence & Margolis 2001). Este argumento, defendido por Chomsky (Chomsky 1965), se basa en ¿cómo puede un niño haber aprendido las reglas sintácticas y gramaticales de un lenguaje con tan pocas instancias? Su conclusión es que hay mecanismos innatos en el ser humano, en concreto, un dispositivo ubicado en el cerebro para la adquisición del lenguaje. Comprobó que la adquisición del lenguaje en niños de corta edad no dependía de la lengua en concreto, y de ahí propuso la existencia de una gramática universal innata en los seres humanos. Su propuesta se concreta en que el lenguaje es algo innato del ser humano, y no aprendido.

El humano clasificador podría verse afectado por esto en su proceso de clasificación, naturalmente. No obstante, el ejemplo que se trabaja en nuestro experimento (timeofday_day) difícilmente sería considerado un concepto innato por algún innatista, y no parece que dependa de la "estructura de la mente". Un experimento que pretendiera hacer aprender a la máquina a partir de imágenes conceptos más atractivos para el innatista (piénsese en conceptos como "negación", "identidad", "conjunción", etc.) sería bien distinto, si es que resultara viable. En todo caso, nuestro experimento plantea el aprendizaje de un concepto de forma empírica, a partir de un conjunto de imágenes de entrenamiento, tal como haríamos, en principio, para enseñar a un niño conceptos como 'rojo' o 'perro'. El procedimiento presupone una versión de la teoría del prototipo (v. supra 3.2, p. 78), ya que es la elaboración de un prototipo precisamente lo que faculta para clasificar con éxito nuevas instancias bajo el concepto aprendido.

La peculiar versión de innatismo respecto a los conceptos defendida en Fodor (Fodor 1984[1975]) se apoya en la idea de que los modelos de aprendizaje de los conceptos tratan a estos como hipótesis, lo cual supone que hay cierta captación previa del concepto por parte de quien supuestamente ha de aprenderlo. Su conclusión es que los conceptos léxicos son innatos. Esta posición radical de Fodor fue modificada tras los debates suscitados por su propuesta (Fodor 2008). Actualmente sigue manteniendo que los modelos de aprendizaje tratan a los conceptos como hipótesis. En este sentido, hay que probar la hipótesis, que ya se tiene, y de ahí se sigue que no hay conceptos que se puedan aprender, ni simples ni complejos. Pero ahora defiende que para este argumento no se requiere que los conceptos

sean innatos. Sugiere que los conceptos, algunos y no todos, se adquieren a través de procesos que son en gran medida biológicos y que no admiten una descripción a nivel psicológico. De esta forma admite la posibilidad de conceptos aprendidos del entorno del individuo. El hecho de que, el sistema conceptual humano sea muy sensible a su entorno cultural, permite la posibilidad de aprender conceptos que nada tienen que ver con los procesos biológicos. El aprendizaje de estos conceptos es un logro cognitivo y no biológico.

Dado que nosotros nos hemos referido aquí a nuestro ordenador como un constructor de hipótesis, el argumento de Fodor tal vez sea relevante aquí, podría pensarse. Se debe hacer constar, no obstante, que esta tesis trata, en el fondo, de un método de clasificación o categorización que debe resolver una máquina, y que debe responder a las expectativas de una mente humana. A fin de cuentas la máquina realiza un proceso de descubrimiento que no es estrictamente un proceso de 'replicación': no 'replica' el concepto que tiene previamente el humano, ya que aunque la base de entrenamiento que se le proporciona tiene clasificadas las imágenes en positivas y negativas, y esa clasificación ha sido hecha por un humano, la máquina elabora su propia hipótesis, y lo que se persigue en el experimento es una eficiencia estadística en la predicción respecto a lo que haría un humano. Podría decirse entonces que el concepto que la máquina aprende (supuestamente) es parasitario del humano, de acuerdo, y que la cuestión relevante en relación al carácter innato o adquirido de los conceptos es cómo ha generado el concepto el humano, y no la máquina. Pues bien, sobre eso nuestro experimento nada nos dice.

En el terreno de la inteligencia artificial, por otro lado, la propuesta innatista vendría a decir que una máquina puede estar diseñada –hardware y software- de tal manera que capture conceptos intrínsecos a su diseño. No podemos entrar en disquisiciones técnicas que nos alejarían del tema que nos ocupa. Baste decir que nuestro experimento asume el carácter aprendido –o al menos "aprendible"- de los conceptos, en la línea de los planteamientos empiristas: aquello que la máquina elabora, sea un concepto en sentido pleno o no, se origina y se construye a partir de imágenes de entrenamiento. Lo que la máquina clasifica son imágenes, y, por tanto, hay una merma importante respecto a la categorización. Pero el programa parte, y elabora, una información visual, aunque en sentido estricto, no vea, en tanto no percibe formas con significado. Este trabajo es llevado a cabo por un hardware de

propósito general gobernado por un software que de ningún modo contiene antes de comenzar la tarea (eso sería aquí el análogo al innatismo en humanos) los conceptos que debe aplicar exitosamente posteriormente. Por tanto, podemos afirmar que nuestra máquina, tanto en su hardware como el software, no potencia "a priori" ningún concepto, y en este sentido hemos de descartar que la máquina tenga conceptos innatos o no aprendidos.

Esta afirmación no contradice el hecho de que debido al hardware y al software implementado pudiera haber 'ventajas' en la captación/elaboración de ciertos conceptos, por ejemplo, aquellos conceptos ligados íntimamente a color y/o textura. Estas 'ventajas' hay que entenderlas respecto al aprendizaje de otros conceptos menos relacionados con el color y/o textura. Por ejemplo, se cometerían menos errores si el concepto a aprender fuera un 'coche rojo' que al capturar un concepto como 'timeofday_day' visto en el experimento. Este hecho puede ser explicado porque las características de bajo nivel (descriptores) codifican aspectos físicos básicos de color y textura, con gran dependencia de un mismo color para uno de los conceptos. (Benavent et al. 2010). No obstante, ¿acaso no serían esos mismos conceptos los que menos problemas de aprendizaje plantean también a los humanos? No podemos sino plantear la pregunta, pues el estudio de cómo aprenden los conceptos los humanos desbordaría con mucho los objetivos del presente trabajo.

3.5 *Conclusiones*

Por lo visto hasta aquí nuestro experimento parece más afín –en el sentido de más fácilmente explicable- a unas que a otras de las distintas posiciones del marco teórico comentadas. Así,

- (1^o) La distinción entre contenido conceptual / no conceptual desde una perspectiva representacional (para identificar los elementos del experimento), se resume en:

  Contenido no conceptual

  - las imágenes, cualquier imagen de la base de datos
  - las matrices HSV y RGB,

  - los descriptores y
  - las características virtuales.

  Contenido candidato a conceptual

  - los centroides,
  - los vectores beta y el límite de equiprobabilidad;

  estos últimos elementos son el contenido aprendido para la tarea de clasificación.

  Hemos visto que nuestro experimento, en general, no supera la "generality constraint" propuesta por Evans para identificar contenido conceptual (v. supra 3.1, p. 70), pero sí la definición de un concepto 'minimalista' propuesto por Camp, con las matizaciones vistas en (v. supra 3.1, p. 71). Además, con el argumento de Roskies, estos elementos aprendidos serían considerados como conceptos incipientes (v. supra 3.1, p. 74).

- (2º) En relación a las diversas alternativas sobre la estructura de los conceptos, la teoría del "prototipo" parece la más adecuada para explicar cómo, a partir de imágenes positivas y negativas extraemos el concepto que buscamos y, posteriormente, clasificamos las instancias según el concepto aprendido. El prototipo es la instancia estándar (el prototipo del concepto pájaro podría ser algo parecido a un gorrión o un mirlo (antes que un loro o un buitre). El centroide utilizado para la elección de imágenes en cierto sentido podría interpretarse como un ideal de la imagen visual que instancia un concepto determinado. Por una parte el centroide no es ninguna imagen existente; por otra, esta imagen virtual integra coherentemente información de todas las instancias positivas en el espacio de representación.

  También podríamos interpretar las imágenes positivas más cercana al centroide como 'las más semejantes' al concepto analizado. En este sentido, sería coherente que la Baseline1, cuyas instancias de entrenamiento positivo han sido tomadas como las más cercanas al centroide son en realidad las más semejantes "virtualmente" –definibles matemáticamente- del concepto, y que

por el hecho de construir el modelo en base a estas imágenes, se espere un mejor rendimiento explicativo y predictivo, según los resultados obtenidos.

Nuestro modelo necesita para calcular la medida de semejanza los vectores beta, y puede prescindir del centroide como prototipo para este cálculo. Por tanto, estos vectores también contienen información coherente del concepto.

En cuanto al centroide construido a partir de las instancias negativas, cuyo papel en el experimento, como ya se ha señalado, es fundamental, cabe distinguir entre el que se obtiene a partir de las imágenes que no pertenecen a la misma familia a la que pertenece el concepto (este es el que utiliza la Baseline1 y Baseline2) y las que sí pertenecen (Baseline3). El primero se apoyaría en un conjunto de imágenes más heterogéneo, mientras que el segundo elegiría aquellas imágenes (negativas) conceptualmente emparentadas con el concepto objeto de investigación. No entraremos aquí en su posible caracterización como "antiprototipos virtuales".

- (3º) En cuanto al estatus ontológico de los conceptos, si el experimento evidencia algún tipo de aprehensión conceptual, entonces los conceptos aunque puedan ser considerados como representaciones mentales –o su equivalente en una computadora–, difícilmente lo serán como sentidos fregeanos, serían más bien como capacidades.

  En este sentido, serían tomados como una suerte de disposiciones para efectuar ciertas tareas, en particular como una habilidad discriminativa. En consecuencia podemos afirmar que un agente conoce -ha aprendido- un determinado concepto si tiene la habilidad de, dada una imagen cualquiera, decidir correctamente si sería adscrita o no a dicho concepto, donde 'correctamente' significa que la computadora coincidiría en su veredicto con el de un humano.

- (4º) A propósito de la cuestión innato/aprendido, el experimento parte de un contexto donde el concepto se genera a partir de información visual (imágenes). El ordenador posee cierta capacidad de cálculo (ese sería el aparato "innato"), que sirve de base a una capacidad/habilidad posterior que consiste en discriminar imágenes. Por tanto, la investigación se plantea, desde su inicio, como una investigación sobre la posibilidad de aprender/generar un concep-

to que, por descontado, el ordenador no posee previamente, con lo que no cabe esperar que el experimento, en principio, pueda aportar consecuencias relevantes al respecto.

En el siguiente capítulo propondremos una alternativa representacional, con matices respecto a las teorías vistas en este capítulo, frente a una alternativa disposicional que sigue siendo la que mejor encaja con las coordenadas de nuestro experimento.

Si buscas resultados distintos, no hagas siempre lo mismo.

Albert Einstein

CAPÍTULO

4

Representaciones vs. disposiciones

En los capítulos anteriores hemos mostrado el experimento de clasificación y tras un análisis crítico (v. supra capítulo 2, p. 5), se procedió a discutir el encuadre de sus elementos más importantes con los diferentes enfoques sobre teorías de conceptos filosóficas y de las ciencias cognitivas (v. supra capítulo 3, p. 61). Los resultados de estas discusiones apuntaban a dos posibles alternativas que podrían aclarar el tema principal de la tesis, esto es, ***si las máquinas pueden poseer conceptos***.

En el presente capítulo intentaremos, en primer lugar, encontrar el encuadre del experimento con una alternativa representacional de conceptos, en concreto con la teoría neo-empirista de conceptos. Seguiremos con críticas a proyectos actuales similares a nuestro experimento. En segundo lugar se discutirá una alternativa, la concepción disposicionalista de los conceptos, que según apuntamos en el capítulo anterior (v. supra capítulo 3.3, p. 93) parecen encajar mejor con las coordenadas del experimento.

4.1 *Una alternativa representacional*

¿Es posible encontrar alguna teoría representacional que muestre que nuestro experimento es capaz de capturar conceptos? Y si es así, ¿qué papel jugarían los elementos de nuestro experimento?, ¿qué restricciones habrá que imponer? Y si no es así, ¿puede nuestro experimento mostrar que se ha aprehendido el concepto sin saber utilizarlo?

En esta sección trataremos sobre el encuadre de nuestro experimento dentro de la teoría neo-empirista de conceptos. Esta teoría resulta atractiva para los trabajos de ingeniería ya que trata con artefactos que pueden reducirse a máquinas. Seguidamente presentaremos críticamente algunos proyectos actuales que trabajan en líneas similares a las de nuestro experimento.

(a) Encuadre neo-empirista del experimento

Las teorías neo-empiristas de conceptos sostienen que los conceptos son copias o combinaciones de copias de representaciones perceptivas. Jesse Prinz, uno de los autores actuales más relevantes en esta línea, defiende el neo-empirismo de conceptos a partir de un argumento de parsimonia, según el cual las representaciones perceptivas son suficientes para dar cuenta de todos los fenómenos que una teoría de conceptos debería explicar, de modo que postular representaciones amodales sería innecesario (Prinz 2002, 2005).

Por otro lado, las teorías neo-empiristas de conceptos afirman que los vehículos de estos son las representaciones perceptivas (RP) o modales (v. infra 105 y p. 109). Y están comprometidas con la afirmación de que los vehículos del pensamiento son reactivaciones de las representaciones perceptivas. Así, Prinz amplia la versión de la máxima de Hume : "Todas nuestras ideas no son más que copias de nuestras impresiones "(Hume, 1748/1975, p. 686), y propone que todas las representaciones conceptuales están codificadas en una modalidad sensorial:

> *"Concepts are couched in representational codes that are specific to our perceptual systems"*
>
> (Prinz 2002 p.119).

Prinz llama a esta afirmación *'la hipótesis de especificidad modal'*. Barsalou (Barsalou 1999) suaviza esta hipótesis admitiendo que "algunas representaciones amo-

dales son necesarias para explicar el funcionamiento de la introspección ".

En este sentido, según Dove (Dove 2009), tanto Barsalou como Prinz están comprometidos con la idea de que los símbolos perceptivos [25] implican simulaciones de la experiencia. Con otras palabras, proponen que nuestra conceptualización de una categoría consiste en simulaciones de las experiencias de los ejemplares percibidos de esa categoría. Tales simulaciones son el resultado de una especie de recreaciones neurofisiológicas[26] (Barsalou 1999).

Nuestro interés se centrará sólo en una parte de la percepción, la correspondiente a la visión bidimensional estática. Donde, las representaciones modales de las imágenes aparecen de forma natural en forma de sus matrices RGB y HSV. Estas matrices, deben su forma a los mecanismos que utiliza la máquina para poder capturar una imagen y, sobre todo mostrar una imagen por pantalla. Por tanto, la codificación está fuertemente relacionada con los mecanismos físicos de la percepción.

Esto es coherente, en principio, con lo que sostiene Prinz, sobre lo que hace que una representación sea modal o perceptiva. Afirma que:

> *"...perceptual representations are representations in dedicated input systems and that dedicated input systems use disparate kinds of mental representations."* (Prinz 2002 p.119).

Dove sostiene, con otras palabras, la definición de representación modal como:

> *"... a representation is modal if it is part of a specific sensory code: that is, if it is contained within a neural system specifically designed through natural selection to detect internal or external objects or events."* (Dove 2009, p. 415).

Ahora argumentaré esta coherencia.

[25] Barsalou identifica los conceptos como símbolos perceptivos. Prinz con proxitipos (v. infra 4.1, p. 111), pero según Prinz (Prinz 2002 p. 152) ambos términos son sinónimos.

[26] El proceso consistiría en, primero, capturar y almacenar la información relativa a la activación neuronal de patrones asociados con la percepción de un objeto o evento, por neuronas conjuntivas, en áreas de asociación vecinas o en zonas de convergencia (Damasio 1989 y Damasio & Damasio 1994) y, segundo, esta información almacenada se utiliza después, en ausencia de entrada perceptiva, para generar una parcial reactivación de las representaciones sensoriales.

Para mantener esta afirmación de coherencia es necesario matizar que Prinz define las representaciones modales pensando sobre todo en humanos, aunque se podría extender también a animales. En cambio, el actor de nuestro experimento es una máquina. En este sentido, vincula el *"ser parte de un código sensorial especifico"* a cómo son los códigos sensoriales específicos en humanos, es decir, que *"esté contenido dentro de un sistema neural diseñado específicamente a través de selección natural para detectar objetos o eventos internos o externos"*, señalando su diseño dentro de las teorías darwinistas de la evolución natural. En nuestro caso, las definiciones i) *"ser parte de un código sensorial especifico"* y ii) *"estar contenido dentro de un sistema neural diseñado específicamente a través de selección natural para detectar objetos o eventos internos o externos"* deben adaptarse a la naturaleza del actor al que se le intentan atribuir estas representaciones modales o perceptivas, a la máquina.

Por esta razón no podemos decir que nuestra máquina tenga ni un código sensorial ni un sistema nervioso específicos, ni mucho menos que este último sea producto de una evolución natural para detectar objetos o eventos internos o externos relacionados con la percepción de ese sentido. Para sostener esta coherencia necesitamos más argumentos.

Parece obvio que los artefactos implicados en el sistema necesario para poder llevar a cabo nuestro experimento han sufrido una *evolución*. Desde los artefactos físicos para capturar las imágenes - cámaras fotográficas -, los soportes de estas imágenes - desde el papel hasta los soportes digitales -, los artefactos para tratar estas imágenes - tarjetas gráficas integradas o no en otros artefactos como los ordenadores -, los artefactos para mostrar las imágenes - pantallas, etc. -, los ordenadores; hasta los artefactos no físicos como el software especifico de cada hardware, el software genérico, y el software especifico de algoritmos de clasificación. Todos han ido evolucionando con el tiempo para poder sobrevivir en una economía de mercado. Este mercado impone unas restricciones evolutivas: la competitividad en esta economía propicia a que sobreviva el diseño que mejores prestaciones proporcione para cumplir con su tarea, pero además se le exige que sea económicamente competitivo.

Es cierto que los artefactos *no pueden evolucionar por ellos mismos* en este ambiente, necesitan *un agente humano* para culminar su evolución, pero este agente

humano está sometido a las leyes de una economía de mercado y además a las leyes de la evolución natural de su especie. Para mejorar su supervivencia, el agente humano necesita mejorar el diseño de sus artefactos en una economía mercado, o, de lo contrario lo que peligran son los ingresos económicos que posibilitan su propia supervivencia.

Por otra parte, la evolución de estos artefactos ha ido imponiendo un código específico a los datos capturados y tratados, y lo ha hecho de acuerdo con la evolución - dentro de los parámetros evolutivos mencionados anteriormente - de estos. Es decir, el formato de los datos está siendo condicionado por la evolución de los artefactos.

Además, la evolución del software específico de clasificación sufre también las restricciones impuestas por la economía de mercado y la evolución natural del agente humano. Por tanto, sobrevivirá el algoritmo que mejor - más barato, más eficiente y con menos consumo temporal - resuelva el problema. Durante mi experiencia en este campo, he probado bastantes algoritmos de clasificación de imágenes, y con el que mejores resultados contrastados he obtenido es con el que se ha mostrado en este experimento, por tanto es el superviviente de un proceso competitivo que seguro no será el definitivo.

Resumiendo, i) "ser parte de un código sensorial especifico" para la máquina de nuestro experimento significa *"ser el formato de los datos que se transmiten a través de un hardware y software específicos para capturar la imagen, un formato condicionado por los elementos participantes en la captura de la imagen"*. ii) "ser un sistema neural diseñado específicamente a través de selección natural para detectar objetos o eventos internos o externos", para nuestro experimento significa *"ser los buses de datos y los algoritmos diseñados específica y directamente por el agente humano, por tanto, sometidos tanto a la evolución natural como a la economía de mercado con el fin de detectar tanto los objetos o eventos internos o externos relacionados con la visión bidimensional de imágenes"*. No obstante, y aunque el vehículo en los mecanismos sensoriales de nuestro experimento - cámara, pantalla y sus circuitos electrónicos - sean estas matrices RGB y HSV, debemos implementar mecanismos de abstracción que posibiliten la tratabilidad computacional de muchas imágenes. En este sentido admitimos cierta reducción de la realidad, sin que suponga excesiva pérdida de identidad, es el proceso de cuantización de los

píxeles de la imagen que, junto con la extracción de patrones de textura conforman los descriptores de la imagen, el vector de descriptores (293 d.). Estos procesos se han detallado ampliamente en secciones anteriores.

Aunque sean distintos los vehículos de percepción, la imagen HSV y el vector de descriptores, lo que sí parece indiscutible es que ambos son modales, es decir, dependientes de la información contenida en imágenes, que un agente humano ha clasificado según su percepción de ellas. Entonces, en tanto que la máquina toma como input una base clasificada por un humano que sí ha percibido las imágenes, a propósito de nuestra máquina puede hablarse de una percepción o sentido visual vicarios, en todo caso.

Si consideramos cómo sería una percepción auditiva en una máquina, los formatos de los datos que vamos registrando y transformando en el ordenador estarán también condicionados por el hardware y software utilizado para su captura y sus primeras transformaciones, y, por esta razón, serán muy distintos a los formatos utilizados en nuestro experimento.

En efecto, hay que tener en cuenta que lo que se codifica en la matriz HSV y en el vector de descriptores es la imagen, y no ningún concepto. Aunque sería discutible la pre-conceptualización en el sentido de que el vector de descriptores puede corresponder a más de una imagen, no es relevante para esta discusión en este momento. Lo relevante, en nuestra opinión, es que la información que se proporciona a la máquina la vehiculan imágenes, y dicha información es relativa a la textura, el brillo y el color de estas, propiedades todas ellas sensoriales, en principio.

Una cuestión que puede ser interesante resaltar aquí es la irreversibilidad de la cuantización, que por una parte imposibilita la transformación de un vector de descriptores en una matriz HSV, y por consiguiente imposibilita su visualización en pantalla. Si la información que se guarda en la memoria - digamos a largo plazo - del ordenador es el vector de descriptores tendremos un grave problema de reactivación de la imagen en pantalla, ya que no podremos recuperar la imagen concreta que permite construir el vector. Aunque en la base de datos con la que la máquina trabaja hay una relación biunívoca entre imágenes y vectores de descripción (cada imagen está emparejada con un vector descriptor y cada vector descriptor, se corresponde con una de las imágenes), el contenido informativo del vector no es suficiente para reproducir exactamente la imagen a partir de la cual se generó.

Es por esta cuestión por lo que considero que, aunque el vector de descriptores cumple con su cometido en aras al aprendizaje, él mismo no puede ser el representante de la imagen en todos los dominios. Por tanto, hay que establecer un vinculo entre la matriz HSV y su vector de descriptores.

Esta última puntualización tendrá su sentido cuando intentemos trabajar a partir de conceptos, ya que parece probado por la neurociencia que cuando se realizan tareas cognitivas que implican activación de conceptos, los patrones de activación cerebral que se activan son, entre otros, los que se activaron en tareas de percepción (Barsalou et al. 2003). Por tanto, y a favor de los neo-empiristas, esto es una evidencia de que los vehículos conceptuales son representaciones perceptivas o copias de estas - y en este sentido, las copias se corresponden con la reactivación de patrones perceptivos -. Es decir, imaginar tendrá el sentido, entre otros, de reconstruir y/o recombinar copias perceptivas. Nuestro experimento no trata de cómo operar con conceptos, pero me ha parecido interesante señalar estas posibles consecuencias de la irreversibilidad de la transformación de cuantización.

Por otra parte, en la definición de Prinz sobre cuándo una representación es perceptiva (v. supra 105), y en su discusión previa (Prinz 2002 pp.110-119), apela a diferentes niveles de representación en la visión, cada uno con diferentes propiedades. Estos determinan la forma de ser representada una imagen, tanto el nivel de 'mapa de bits' como el que generan las distintas transformaciones y cuantizaciones, pueden considerarse como diferentes niveles de representación, pero siendo todas representaciones perceptivas. Otra cosa es que comúnmente, los humanos, asociemos una representación perceptiva a unas formas visuales determinadas que puedan descomponerse en objetos con significado en un nivel más elaborado o más complejo de representación. Parece, entonces, plausible admitir tanto la matriz HSV como el vector de descriptores como representaciones perceptivas modales, cuyos formatos son dependientes de la información perceptible contenida en las imágenes, incluso si trabajáramos con las características virtuales, o con los centroides, o con los conjuntos de imágenes positivas y negativas, el formato de todos ellos, está siempre vinculado a esta información perceptible.

Si consideramos al vector beta, junto con el límite de la equiprobabilidad, y el símbolo lingüístico como integrantes únicos del concepto, este no va a cumplir con muchas de las exigencias que se le deben hacer a un concepto.

Para los neo-empiristas, como ya hemos visto, cuando se realiza una tarea cognitiva que necesita activar conceptos, se activan patrones perceptivos, entre otros. ¿Quiere decir esto que se necesita una imagen mental, en nuestro caso visual, para representar el concepto? No necesariamente, pero sí que parece importante que debamos ser capaces de reactivar los circuitos neuronales para activar o crear una copia mental. Si debemos activar una copia, esta debería estar almacenada en - la memoria a largo plazo - del ordenador y estar asociada al concepto.

En nuestro experimento, el elemento más representativo del concepto que se pretende capturar es el centroide de imágenes positivas (con todos los matices vistos hasta aquí en las secciones previas), ya que él es el centro euclídeo de las instancias positivas en el hiper-espacio de los descriptores. Pero, este elemento está representado en el vehículo del vector de descriptores, y no tiene ninguna matriz HSV o RGB asociada. Y, tal como sabemos, no podemos recrear una imagen RGB o HSV a partir de su vector de descriptores, por la irreversibilidad del proceso de cuantización discutido repetidamente en esta tesis. Por tanto, a partir del centroide no podemos ni reactivar - porque no hay ninguna imagen RGB o HSV asociada al centroide guardada -, ni recrear - por el problema de la irreversibilidad -, estaríamos ante un concepto visual 'ciego', es decir, sin conexión directa con ninguna imagen particular, lo que resulta un tanto contradictorio.

Descartado el centroide como representante visual del concepto, ¿qué elementos nos quedan? Los mejores candidatos parecen las imágenes positivas que han servido de base para entrenar el concepto. En este sentido, sí que tenemos imagen visual, y además vector de descriptores y características virtuales. Pero, ¿son necesarias todas las imágenes o podemos conformarnos con la imagen más próxima al centroide? La respuesta que demos a esta pregunta repercutirá en la 'capacidad discursiva' del concepto. En este sentido, si solamente tomamos una imagen, cuando elaboremos un pensamiento con este concepto, su imagen representante aparecerá en escena y no admitirá ninguna interpretación discursiva en el entorno del escenario del pensamiento. Si tomamos todas las imágenes positivas, ante este mismo pensamiento, podemos, en función del escenario donde actúa el concepto, colocar la imagen más adecuada. Esta segunda alternativa admite la característica discursiva del concepto. El cómo lo hará escapa de la presente tesis, pero no sería descabellado que lo encontrara utilizando redes semánticas con las características

virtuales, por ejemplo, o con los mismos conceptos simbolizados por palabras de un lenguaje.

Un apunte más sobre el encuadre del experimento en las teorías neo-empiristas nos lo puede proporcionar la forma en la que elegimos los conjuntos de las imágenes de entrenamiento. Prinz en (Prinz 2002) propone la teoría de los "proxitipos" [proxytypes]. Prinz afirma que los conceptos son proxitipos y los define como conjuntos de representaciones perceptivas que se activan en la memoria de trabajo para detectar categorías, es decir, los conceptos son representaciones mentales de las categorías. Con palabras de Prinz:

> *"... concepts are mental representations of categories that are or can be activated in working memory. I call these representations "proxytypes", because they stand in as proxies for the categories they represent."* (Prinz 2002 p. 149)

Estos conjuntos varían en función del contexto. En este sentido, aunque tengamos por ejemplo el concepto 'perro', no usaremos el mismo conjunto de representaciones perceptivas para buscar un "chihuahua" que un "bulldog". Aquí surge un problema al identificar proxitipos y conceptos, es decir, ¿se usan dos proxitipos distintos para buscar un chihuahua y un bulldog, del mismo concepto 'perro'? ¿O se trata de dos conceptos distintos? Prinz matiza su definición de proxitipos admitiendo un solapamiento o superposición entre ellos, es decir, el 'chihuahua' tendrá su proxitipo y el 'bulldog' el suyo, que estarán solapados en la memoria a largo plazo. Admite que el contexto donde vaya a emplearse determinará qué proxitipo utilizará la memoria de trabajo. Además, junto con Barsalou (que considera los proxitipos como 'símbolos perceptivos'), afirman que si los conceptos son los proxitipos, el pensamiento es un proceso de simulación (Barsalou 1999), (Prinz 2002, p. 150). Prinz va más allá al sostener que una copia particular de un proxitipo (un 'token' de un objeto, por ejemplo), es generalmente equivalente a entrar en un estado perceptivo de este objeto donde se experimenta lo que representa. Podemos simular la manipulación de este objeto real, manipulando los proxitipos cuando este no está. (Prinz 2002, p. 150-151)

Aquí subyace un mecanismo de búsqueda en la base de proxitipos, para seleccionar el proxitipo adecuado según el contexto, que necesita también ser explicado.

Por otra parte, plantea como solución a la categorización más general (cómo encontrar el proxitipo de 'perro', por ejemplo) apelando a la frecuencia con que las características aparecen en esta familia. Es decir, las características que más se repiten forman un 'proxitipo por defecto', que es del que se extraerá el token (una copia particular del concepto) cuando no importa el contexto donde va a ser utilizado (Prinz 2002 pp. 148-157).

En nuestro experimento, al igual que Prinz distingue dos proxitipos diferentes, uno para "chihuahua" y uno para "bulldog", de la misma familia "perro", también hemos adoptado el criterio de que son dos conceptos distintos. Véase que el concepto "timeofday_day" es distinto al "timeofday_night", aunque ambos pertenezcan a la familia "timeofday". En el experimento lo hemos resuelto como dos conceptos distintos. Prinz resuelve la familia 'perro' como un solapamiento de características de sus integrantes. En el experimento hemos dejado la aprehensión de la familia para una red semántica posterior que no se ha tratado, pero que bien podría implementarse como la unión de las instancias que caen bajo los conceptos pertenecientes a cualquier concepto de la familia.

Esta identificación con los proxitipos abre una profunda discusión con el modo de elegir las imágenes positivas y negativas, y por consiguiente con la relatividad del significado real del concepto, que desemboca en su no universalidad. En este sentido, cuando intentamos definir un concepto es muy importante tener en cuenta en qué escenario nos estamos moviendo, frente a qué otras cosas queremos que tenga significado el concepto, qué **no** es el concepto. Es ahí donde tiene sentido la elección de las imágenes positivas, pero sobre todo las negativas. Como se vio en el experimento, una de las opciones de elección de las imágenes negativas en la opción automática distinguía entre elegir el centroide de todas las imágenes que no contuvieran el concepto a tratar ni pertenecieran a la familia (baseline 1 y baseline 2), y otra como el centroide de todas la imágenes que no contuvieran el concepto pero que pertenecieran a la misma familia (baseline 3). Este problema no está resuelto en el neo-empirismo ni tampoco en el experimento realizado en la presente tesis, ya que las diferencias obtenidas entre baseline 2 y baseline 3 en el experimento no muestran resultados concluyentes (v. supra 2.2.3, p. 42).

Para explicar las simulaciones de las experiencias previas (v. supra p. 105) en nuestro experimento, admitamos que, el hecho de entrenar la máquina ha produci-

do como resultado el vector beta, y este, de alguna manera aglutina las simulaciones de la experiencia del concepto aprendido. En concreto, las experiencias de las imágenes de la base de entrenamiento utilizadas para obtener el vector beta. De tal manera que este vector tiene contenido que es capaz de identificar instancias del concepto entrenado. Podemos identificar este vector con los patrones asociados al concepto aprendido durante la fase de entrenamiento, pero sin que pueda recrear una imagen del concepto. [27].

Dove siguiendo a Barsalou, afirma que el sistema conceptual debe entenderse en términos de 'simuladores', entendidos como un sistema distribuido que abarca áreas asociativas y perceptivas que generan simulaciones, y de sus 'simulaciones', generadas por aquéllas y entendidas como el resultado de una especie de recreaciones neurofisiológicas (Barsalou 1999).

El simulador de nuestro experimento es difícil de encontrar, aunque podría identificarse con el mecanismo que genere un vector de descriptores de una imagen que pertenezca al concepto -que nuestro experimento no ha tratado-, incluyendo el vector beta del concepto que se pretende simular como parte del concepto; también la simulación es difícil de encontrar si necesitamos que esta sea una imagen, porque aunque tuviéramos el vector de descriptores de esa imagen, volveríamos a chocar con la irreversibilidad de la transformación.

Por otra parte si *poseer un concepto es tener la habilidad o capacidad de generar unas apropiadas representaciones perceptivas del concepto en una situación dada* (Barsalou 1999, 2003), entonces, con esta definición de posesión de conceptos, considerando sólo las imágenes como representaciones perceptivas, podemos afirmar categóricamente que nuestro experimento muestra que '*nuestra máquina NO puede poseer conceptos*', y esta afirmación se basa en la imposibilidad de generar representaciones perceptivas al uso, por el problema de la irreversibilidad.

Ahora bien, según Barsalou y Prinz, la conceptualización de una categoria consiste en las simulaciones de ejemplares de esa categoría, y la categorización se puede explicar en términos del grado de ajuste entre la percepción real y la simulación (Prinz 2002 cap. 6.3). Veamos, de algún modo, la simulación se incorpora en

[27]Como ya hemos dicho en varias ocasiones, no podemos a partir del vector de descriptores de una imagen, y menos de los parámetros del vector beta de un concepto, recrear la imagen original ni la imagen del concepto por el problema de la irreversibilidad. La simulación ha de entenderse como la recreación del vector beta ante una imagen de entrada que se pretende clasificar.

el resultado de la aplicación de la función de regresión logística de un determinado número de instancias positivas y negativas para entrenar el concepto a simular, o sea, el vector beta. En tal caso, el grado de ajuste entre la percepción real, representada por su vector de descriptores, y el concepto, representado por el vector beta, proporciona el grado de similitud de la instancia, que es el objetivo de la categorización que propone Barsalou y Prinz. A efectos prácticos, se debe exigir que sea una *representación perceptiva*, independientemente de que provenga o no de una simulación perceptivamente visible. Y representaciones perceptivas las tenemos en las representaciones como el vector de descriptores de las imágenes de entrenamiento, al menos ahí, y si consideramos como experiencia perceptiva el vector beta elaborado a partir de las instancias de entrenamiento, este puede ser un perfecto candidato a representar la simulación de la experiencia perceptiva. También es fácil encontrar representaciones en forma de vector de descriptores al ajustar vectores que cumplan el requisito de dar positivo para un determinado concepto, y en este sentido, la productividad de simulaciones se dispara.

Admitamos de momento que el vector beta de un concepto representa la simulación de las experiencias perceptivas previas sobre ese concepto. Admitamos que el proceso de categorización consiste en generar este vector beta según los procedimientos vistos en esta tesis, y que por tanto, tiene en cuenta las experiencias perceptivas de las imágenes de entrenamiento para su construcción. Admitamos que este proceso de categorización se lleva a cabo dinamicamente, es decir, en una especie de memoria de trabajo -como propone Prinz- y que tiene en cuenta para la construcción del vector beta del concepto, las instancias de entrenamiento acordes al entorno donde pretende que tenga sentido el concepto. Ante estos presupuestos, es compatible nuestro vector beta con las exigencias que Prinz y Barsalou imponen a las representaciones perceptivas para ser representantes de conceptos, 'símbolos perceptuales o proxitipos' según estos autores [28].

Nótese que el sentido que le hemos dado a las simulaciones es el mismo que sostienen Prinz y Barsalou sobre la capacidad de 'imaginar' a partir de manipulaciones de copias de la percepción. En ningún caso estamos hablando de simuladores y simulaciones que son llevadas a cabo por mecanismos informáticos dedicados

[28] Aun así queda por explicar la composicionalidad de conceptos y su simulación, es decir, la generación de nuevos conceptos a partir de conceptos previos.

a campos de entretenimiento y/o formación. Sobre estos últimos, queda claro que comercialmente las máquinas simuladoras se han impuesto socialmente, y no cabe preguntarse si la máquina tiene la capacidad de realizar simulaciones, de hecho, se pueden ver como simuladores. Y estos simuladores van incluso más allá de las exigencias de Prinz y Barsalou en el sentido de que no solamente realizan una simulación 'psicológica' o 'mental' sino que además, la hacen visible para los humanos, y en algunos casos, reproducen el entorno físico de la simulación. Piénsese, por ejemplo en una máquina simuladora de vuelo para entrenar pilotos humanos. Si este fuera el caso, el hecho de realizar simulaciones nos llevaría inmediatamente a afirmar que la máquina sí tiene sistema conceptual. Nuestro razonamiento se centra exclusivamente en el experimento, y en este sentido hay que analizar la capacidad de realizar las simulaciones siguiendo el protocolo que hemos establecido, es decir, partiendo de la capacidad de clasificación de instancias bajo conceptos, ser capaz de 'imaginar' o simular escenarios donde aparezcan de una forma determinada estos conceptos. La carga de la simulación recae sobre el vector beta como recreación de la experiencia previa sobre las instancias de entrenamiento. Nuestra posición es que, con todo lo visto hasta ahora, no es posible atribuir simulaciones, en el sentido que sostienen Prinz y Barsalou, a nuestro experimento.

Además, cabría preguntarnos si el neo-empirismo puede prescindir de representaciones amodales. Muchos críticos, e incluso no críticos, del neo-empirismo responden con argumentos poderosos que muestran que no se puede prescindir de ellos (Dove 2009). Sobre todo basan sus argumentos en las zonas de convergencia de los sentidos. Es decir, en estas zonas, tanto las representaciones modales de un sentido, por ejemplo el visual, como de otro sentido, por ejemplo el auditivo, se deben acoplar entre ellas para formar un concepto que suponga una combinación de ambas percepciones. Y para combinarse es necesario un código que no dependa del sentido de la percepción, ni el auditivo, ni el visual. La cuestión es si se puede dar cuenta de este tipo de combinaciones sin apelar a representaciones que trasciendan el nivel sensorial ligado a uno u otro sentido, esto es, representaciones amodales.

En el experimento, son el vector beta, el límite de la equiprobabilidad, y el/los símbolos lingüísticos utilizados para hacer referencia al concepto, y no a la imagen los que debemos examinar con más detenimiento. Aunque el formato del vector beta depende de la función de distribución que hemos usado - el número de dimen-

siones del hiperespacio de descriptores más uno, por ejemplo -, podríamos utilizar también esta función de distribución para calcular la probabilidad de pertenencia a un concepto desde otro sentido de la percepción - por ejemplo el auditivo -, en cuyo caso, los vectores de descriptores se habrían formado de manipular los parámetros específicos del sonido, con sus propios formatos de datos, pero seguramente con distintas dimensiones de las que hemos usado para las imágenes. La salida de la función de probabilidad sí tiene el mismo formato para cualquier sentido, es decir, genera un valor entre 0 y 1 para cada 'sentido' en función de la similitud con el concepto que se esté considerando. Además, el límite de equiprobabilidad proporcionará una salida dicotómica (si, no) sobre la instanciación de un determinado concepto para cada sentido. En definitiva, la salida de la función de probabilidad de cada sentido y la comparación con su propio límite de equiprobabilidad sí que parece que ayude al entendimiento entre los diferentes sentidos, empiezan a hablar el mismo lenguaje. Por ejemplo, según el sentido visual, de una imagen determinada encontramos a la salida de la función de probabilidad y comparada con su límite de equiprobabilidad, que tiene el concepto gato, y, por otra parte, del ladrido que ha escuchado el sentido auditivo al mismo tiempo que se le presentaba la imagen, se llega a la conclusión que no es un gato, ahora sí que se pueden entender los símbolos perceptivos de los dos sentidos. Entre ambos pueden formar el concepto de 'gato que ladra', o, 'perro con apariencia de gato', o, 'he mirado mal', o 'he oído mal', etc, juicios que serían imposibles si solamente consideráramos los vectores beta de ambos sentidos.

En principio esta crítica no afecta a nuestro experimento ya que solamente se ha tratado un sentido, la visión bidimensional y estática. Pero, como ya venimos apuntando a lo largo de esta tesis, la capacidad de realizar inferencias, de explicar el contenido del concepto, y de aprender desde otros mecanismos distintos a la percepción -por medio de lenguaje, por ejemplo-, puede suponer utilizar simbolismo y salir del conexionismo en el que se ha concebido el experimento – entiéndase este simbolismo para construir redes semánticas, etc. a partir de los símbolos o etiquetas de los conceptos-. Este tema escapa de la discusión de la presente tesis.

Veremos a continuación unas críticas generales a esta teoría. En primer lugar, la pertinencia de identificar los 'símbolos perceptivos o proxitipos', cuando se simulan en la memoria de trabajo con las áreas cerebrales que se activan en la percep-

ción externa. En segundo lugar, el problema que tiene esta teoría con los conceptos abstractos, aunque para esta tesis no es relevante.

Primero: El hecho que las simulaciones son el resultado de una especie de recreaciones neurofisiológicas (v. supra p. 105) se sostiene a partir de las evidencias de estudios neurocientíficos que muestran que las regiones de percepción del cerebro se activan durante tareas cognitivas como la categorización y la inferencia. Pero esta evidencia se queda corta, ya que como sostiene Weiskopf en (Weiskopf 2007), durante las tareas conceptuales, la estructura causal del cerebro produce actividad generalizada en ambos sistemas, perceptivo y no perceptivo. Por tanto, concluye que las representaciones perceptivas no pueden señalarse como el único vehículo del pensamiento conceptual, y que hay que considerar representaciones modales y amodales. Weiskopf ve el empirismo conceptual contemporáneo como una tesis sobre la naturaleza de los vehículos del pensamiento.

Segundo: También Machery examina críticamente el resurgimiento del neo-empirismo en (Machery 2006) . A partir de dos tesis que parecen sostener los neo-empiristas, a las que llama dogmas del neo-empirismo:

- (1) El conocimiento que se almacena en un concepto se codifica en varios sistemas de representación de percepción.
- (2) El tratamiento conceptual implica recrear algunos estados de percepción y la manipulación de estos estados de percepción.

La tesis 1 trata sobre el vehículo o la forma de nuestro conocimiento conceptual, cómo codificamos nuestro conocimiento conceptual; y la tesis 2 se refiere a la naturaleza de nuestros procesos cognitivos - categorización, inducción, deducción, etc.-

Machery sostiene que el neo-empirismo está mal equipado para hacer frente a los conceptos y proposiciones abstractas, y además admite que en los procesos cognitivos complejos son compatibles las representaciones modales y amodales.

Concluyendo, aunque podamos acomodar los elementos del experimento expuesto en el capítulo 2 (v. supra 2.3, p. 43), no es suficiente para poder afirmar que estemos ante la captación de un concepto tal como define esta teoría neoempirista. Por una parte, sí podemos encontrar representaciones perceptivas en vehículos amodales. También seríamos capaces de construir una especie de 'proxitipo'con

el vector beta, el límite de equiprobabilidad, y necesitaríamos además el centroide o la imagen positiva más próxima. También podemos explicar la categorización mediante la función de similitud. Pero por otra parte, no podemos generar simulaciones compatibles con todos los vehículos de la percepción vistos; y tampoco podemos realizar recombinaciones de conceptos y después simularlos.

Seguidamente se expondrá breve y críticamente proyectos actuales que siguen algunas de las directrices del experimento expuesto en el capítulo 2, p. 5.

(b) Proyectos similares al experimento

Existen en el mercado modelos en desarrollo que describen textualmente imágenes bidimensionales, por ejemplo, el NeuralTalk2[29] que utiliza los algoritmos matemáticos del paquete Torch[30]; o el "Google's Brain-Inspired Software[31]".

En ambos casos, tras una fase de entrenamiento, el programa intenta describir los objetos que contiene una imagen. Su función consiste básicamente en localizar objetos en la imagen -previamente aprendidos en la fase de entrenamiento- y expresarlos en un lenguaje humano. Ambos necesitan conjuntos de entrenamiento de imágenes que contengan sólo un objeto, que será el objeto a entrenar. Se entrena cada objeto por separado. Los parámetros de las imágenes se extraen de características visuales. Posteriormente, tras el entrenamiento, y con ayuda de redes semánticas de lenguaje humano, se describen las imágenes objetivo. En concreto, definen un escenario con posibles relaciones entre los objetos encontrados.

Estos modelos tienen similitudes con nuestro experimento, en el sentido de que utilizan la información visual y no textual para entrenar un dispositivo localizador e identificador de objetos dentro de una imagen. Pero se diferencian en que estos objetos son muy específicos y altamente imaginables; además las imágenes de entrenamiento que se suministran contienen sólo este objeto con mucha nitidez. De hecho, en nuestro experimento, primero, los conceptos que se intentan capturar son conceptos de difícil concreción en una imagen -p.e. timeofday_day-, y, segundo, las imágenes de entrenamiento pueden contener escenarios con multitud de conceptos juntos.

Por otra parte, aunque pueda dar la impresión de que en la descripción textual

[29] https://github.com/karpathy/neuraltalk2 consultado en abril de 2016

[30] http://torch.ch/ consultado en abril de 2016

[31] https://www.technologyreview.com/s/532666/googles-brain-inspired-software-describes-what-it-sees-in-complex-images/ consultado en abril de 2016

de la imagen que realizan estos proyectos se vislumbre una composicionalidad de conceptos, en realidad, lo que se ha capturado son los objetos que componen la imagen y, 'a posteriori' una red semántica léxica se ha encargado de componer el contenido textual. No se prueba que puedan recombinarse sistemáticamente conceptos y posteriormente simularlos. En este sentido, estos sistemas también adolecen de la irreversibilidad en este nivel, ya que a partir de una descripción textual, no son capaces de mostrar una imagen que recoja este escenario.

Estos sistemas se apoyan con redes semánticas externas. Aunque la dimensión semántica no pretende ser incorporada en nuestro experimento, y por tanto, no es objeto de análisis 'per se' en esta tesis, haré dos comentarios breves:

- *Primero:* Tal como vimos en el apartado anterior (v. supra 4.1, p. 115) puede surgir un problema de entendimiento conceptual en las zonas de confluencia de las percepciones, pero sobre todo para representar conceptos abstractos. En este sentido, el articulo de Dove, (Dove 2009), al mismo tiempo que sostiene que los argumentos en apoyo de los símbolos perceptivos son mucho más sólidos con respecto a conceptos concretos – altamente capaces de formar imágenes- frente a conceptos abstractos, también intenta ser coherente con la existencia de una lengua o lenguas del pensamiento (Fodor 1984[1975]) y con las posiciones que postulan representaciones amodales que no se identifican con ningún lenguaje natural. Su propuesta es que el sistema conceptual humano se caracteriza por una división representacional del trabajo en representaciones modales y amodales que manejan diferentes aspectos de nuestros conceptos.

  Si admitimos la llamada 'teoría dual del código' (Paivio 1987, citado en Dove 2009), existen dos sistemas semánticos, uno apoyado por representaciones lingüísticas y otro con el apoyo de las representaciones perceptivas. Los efectos de la imaginabilidad, -definida como la facilidad con que una palabra da lugar a una imagen mental sensorio-motora- pueden ser explicados en términos de la mayor disponibilidad de información codificada perceptivamente. En este sentido, palabras con baja imaginabilidad son asociadas principalmente con las representaciones verbales mientras que palabras altamente capaces de formar imágenes se asocian tanto con representaciones

lingüísticas como perceptivas.

Entonces, si los conceptos se codifican de forma perceptiva y de forma amodal, ¿por qué no considerar una red semántica léxica asociada al concepto? Las ventajas son obvias, cumplirían las tareas explicativas e inferenciales. Se codificaría simbólicamente y enlazaría bien con los artefactos simbólicos de inteligencia artificial -arboles de decisión, reglas, sistemas basados en el conocimiento, etc.-. Este planteamiento es el que han seguido los ejemplos de proyectos expuestos anteriormente, pero como ya hemos señalado, el estudio de este punto escapa del alcance de esta tesis.

- *Segundo:* Centrándonos en el problema de clasificación del autómata, nuestro grupo de investigación UNED-UV ha publicado en repetidas ocasiones los resultados de distintos experimentos en los que se puede observar que los enfoques simbólicos, es decir, clasificación a partir de las etiquetas textuales de las imágenes, es tan eficiente o más que el enfoque conexionista visual que hemos visto en el experimento de la presente tesis (Benavent et al. 2010; Benavent et al. 2012; Benavent et al. 2013).

  Además, se ha aumentado notablemente la eficiencia cuando se ha optado por un enfoque mixto, es decir, por una parte se ha trabajado simbólicamente con las etiquetas textuales y por otra parte con la información visual de las imágenes. La forma de "mezclar" los resultados de estos procesos tiene también mucha repercusión en la eficiencia de la clasificación. Según nuestras investigaciones, el mejor método de fusión ha sido utilizar como filtro primario la clasificación textual, y después realizar la fusión con los resultados del procesamiento visual -este procesamiento es similar al expuesto en el experimento de la presente tesis-.

  Desafortunadamente, las imágenes reales que percibimos no están etiquetadas textualmente en ningún lenguaje, y en este sentido, aunque para los fines perseguidos por los promotores de los distintos concursos, y sobre todo para el problema de clasificación de imágenes guardadas en bases de datos con información textual, se puede afirmar que la incorporación del procesamiento simbólico es muy valioso, para la clasificación de imágenes procedentes del mundo real no es útil.

Aun así, estos resultados podrían servir como argumento para sostener el enfoque dual que propone entre otros Dove (Dove 2009) y que se ha discutido en el apartado anterior. En este sentido, y aunque lo que se propuso era la creación de redes semánticas a partir de los símbolos obtenidos del proceso de clasificación, considerando siempre el arranque desde la percepción exterior, podrían plantearse procesos de introspección a partir de imágenes almacenadas y etiquetadas en memoria a largo plazo, donde quizá pueda ser útil. Estos planteamientos escapan del alcance de la presente tesis que se centra solamente en los resultados del experimento expuesto en el capítulo 2.

4.2 *Una alternativa disposicional*

Uno de los primeros métodos para examinar la posesión de conceptos en animales surgió de una serie de experimentos realizados con palomas. Se las entrenaba para que detectaran un objeto (por ejemplo un 'árbol'), en una imagen. El conjunto de entrenamiento consistía en imágenes que tenían el objeto y otras que no lo tenían. Tras el entrenamiento, las palomas picaban sobre las imágenes del conjunto de entrenamiento que tenía el objeto, y no picaban sobre las que no lo tenían. Posteriormente fueron capaces de generalizar esta habilidad a nuevos conjuntos de imágenes. Se sugirió que esta capacidad de clasificación demostraba que la paloma tenía un concepto del objeto (Herrnstein 1979).

Esta posición fue criticada como insuficiente para demostrar la posesión de conceptos con el argumento de que, los humanos pueden discriminar un objeto sin poseer su concepto:

> *"It is possible to teach a human being to sort distributors from other parts of car engines based on a family resemblance between shapes of distributors. But this ability would not be enough for us to want to say that the person has the concept of a distributor"* (Allen & Hauser 1996, p. 51).

Este argumento es discutible. Puede que no se demuestre que se tenga el concepto completamente aprendido por el hecho de saber discriminarlo, pero al menos el concepto visual sí se tiene. Además, ¿tiene el mismo conocimiento de esta pieza

un ingeniero mecánico que un conductor novel? Seguro que encontramos dos conceptos diferentes de un mismo referente. Y si es así, ¿no se podría hablar de grados en la posesión del concepto?

Además, estos autores proponen que para poseer un concepto se necesita poder responder de forma distinta ante un mismo estímulo. Es decir, ante un árbol, por ejemplo, responder cobijándose o comiendo sus frutos, en función de los factores ambientales o personales. Pero la tarea de clasificación exige responder de la misma forma ante un mismo estímulo.

Allen presenta una estrategia general para poder atribuir conceptos a animales:

> *"An organism O may reasonably be attributed a concept of X (e.g. TREE) whenever:*
>
> - *(i) O systematically discriminates some Xs from some non-Xs; and*
> - *(ii) O is capable of detecting some of its own discrimination errors between Xs and non-Xs; and*
> - *(iii) O is capable of learning to better discriminate Xs from non-Xs as a consequence of its capacity"*
>
> (Allen 1999, p. 37).

Nuestro experimento muestra la capacidad discriminativa (i), y puede incluso re-aprender (iii) tal como se propuso en el capítulo 2 (v. supra 2.3.5, p. 48), pero no es capaz de detectar sus propios errores, fuera de los conjuntos de entrenamiento, sin ayuda externa.

Si admitimos que es posible una ayuda externa, es decir, que nos señala una instancia mal clasificada, en este escenario contrafactual, tenemos abiertas dos posibilidades:

- *Primera*: siguiendo los procedimientos de reajuste expuestos en (v. supra 2.3.5, p. 48), es posible modificar el límite de equiprobabilidad y, en consecuencia, detectar instancias mal clasificadas anteriormente, con lo que cumpliría el requisito (ii) y reclasificarlas adecuadamente, con lo que cumpliría el requisito (iii).

- *Segunda*: añadir la información de la ayuda externa a la base de datos de entrenamiento y volver a realizar el proceso de obtención de los vectores beta. Aprendidos estos nuevos vectores, cumpliríamos el requisito (iii), detectaríamos los errores de la anterior clasificación y reclasificaríamos, cumpliríamos el requisito (ii).

Quizá el requisito más exigente, incluso para humanos, sea el (ii). Además cabría preguntarse si este es posible con solo un sentido. En este caso, ¿cómo detectamos nuestros errores sin ayuda externa?, o ¿cómo detectamos nuestros errores sin un aprendizaje previo?

Si disponemos de más de un sentido, tal como ocurre en animales y humanos, y aprovechando las posibles contradicciones entre ellos ante una misma entrada, tal como se comentó anteriormente, (v supra capítulo 4.1, p. 116), sí que podemos arbitrar mecanismos que determinen la veracidad del concepto percibido. A partir de esta supuesta veracidad, sí se puede detectar un error en la percepción de un sentido, y, consecuentemente, reaprender el concepto almacenado en el sentido de la percepción que ha 'fallado'. Por ejemplo, ante 'una vista de gato que ladra y huele a perro', se puede construir un mecanismo que identifique este conjunto de percepciones como una instancia de perro atípica, y determine un reaprendizaje desde el sentido visual del concepto perro, añadiendo la nueva instancia visual (aunque visualmente parezca un gato) en la base de entrenamiento positiva para el concepto visual perro. En estos supuestos, sí se cumplen las exigencias de Allen.

Pero, si solo disponemos de un sentido, tal como se ha planteado en el experimento, la situación es distinta. En este caso, para aprender de los errores, primero hay que reconocerlos, y esto, sin intervención externa parece difícil. Si se reconocen errores sin ayuda externa, sus causas pueden ser:

- *Primera*: porque se ha reaprendido el concepto. No se pueden reconocer errores si seguimos en el mismo paradigma y este se aplica correctamente. Pero según Allen, reaprender debe ser la consecuencia del reconocimiento de errores, no la causa.

- *Segunda*: porque se ha aplicado mal la mecánica de clasificación del concepto. En este caso, si tras una revisión posterior de la misma instancia aplicando correctamente la mecánica de clasificación, se detecta el error, se puede

corregir la clasificación previa, pero no se modifica el concepto y no se ha aprendido nada.

- *Tercera*: porque los datos de entrada y/o sus primeras transformaciones son erroneas. Entonces, aunque se aplique correctamente la mecánica de clasificación del concepto, puede ser determinante en una mala clasificación. Tras otro proceso de clasificación de la misma instancia sin errores en las entradas ni sus primeros procesos de transformación, se puede detectar el error. Pero como no se modifican los parámetros que definen el concepto, no se ha aprendido nada. Además, ¿seríamos capaces de identificar como la misma instancia a ambas? Entendemos que no, ya que sus vectores de descriptores podrían ser distintos.

Por tanto, y siguiendo en el supuesto de un solo sentido, si se da (ii) sin que antes haya un reaprendizaje (iii), es decir, se reconoce un error no derivado de un reaprendizaje del concepto, sus causas no apuntan a un error en el concepto, sino a errores en la entrada de datos y/o a la mecánica de clasificación. Si tras reconocer el error se aprende a evitar sus causas, o corregirlas si se producen, en la entrada de datos y/o en la aplicación de la mecánica de clasificación, cumpliría el requisito (iii) sin necesidad de reaprender el concepto. Es decir, aprender a discriminar mejor. Pero estas posibles soluciones no se han tratado en el experimento.

Además, aunque nuestro experimento depende de un solo sentido, entendemos que de modo contrafactual (si se hubieran implementado más sentidos), sí sería posible detectar algunos de los propios errores. Y como consecuencia, se puede reaprender el concepto del sentido que ha fallado, añadiendo esta instancia a la base de entrenamiento y volviendo a generar el vector beta. Un mecanismo similar al expuesto en (v supra capítulo 4.2, p. 123). Este reaprendizaje puede ayudar a clasificar mejor las instancias en un futuro.

En definitiva, con todos los matices introducidos a la estrategia general de Allen, podemos admitir que nuestro experimento muestra que podemos atribuir razonablemente los conceptos aprendidos a la máquina, al menos de forma contrafactual como se ha comentado en los párrafos anteriores.

Por otra parte, y dejando la posesión de conceptos por animales, filósofos como Sir Anthony F. Kenny defienden los conceptos como una capacidad/habilidad

(Kenny 2010). Este autor propone la distinción entre mente, cerebro y conceptos en seres humanos atribuyendo las capacidades a la mente, los conceptos como habilidades donde se manifiestan las capacidades, y el cerebro como vehículo para ejecutar las capacidades mentales.

En concreto define los conceptos como:

> *"We may use 'concept' as a term for the specific abilities that are particular exercises of the universal capacity that is the mind."* (Kenny 2010, p. 105-106)

Atribuye a la mente la *'capacidad'* de adquirir habilidades, con lo cual sitúa la mente como un concepto más según su propia definición, pero la nombra como *'capacidad'*. Esta capacidad, que es intrínseca a un humano, la mente, es distinta y única para cada individuo. En este sentido, la *'capacidad'* de aprender la habilidad de clasificar imágenes bajo conceptos léxicos que propone nuestro experimento, si el agente fuera humano, este autor lo consideraría como una *'capacidad'*, y por tanto, como una tarea propia de una mente. No podemos usar este argumento para atribuir mente a un programa de ordenador, en el fondo, el diseñador del programa sigue siendo un humano, y la *'capacidad'* de diseñar el programa que pueda, o no, tener la *'capacidad'* de aprender la habilidad de clasificar imágenes bajo conceptos léxicos sigue estando en el diseñador del programa. Aún así, ¿podríamos hablar de mente humana y mente mecánica o es siempre reducible a mente humana?, dejamos abierta esta cuestión.

Según el autor citado anteriormente, tanto las capacidades como las habilidades son individuales y además, las habilidades están determinadas por su ejercicio. Es decir, es un individuo quien posee una capacidad y desarrolla una habilidad en un determinado ejercicio. La habilidad está determinada por su ejercicio, pudiendo ser distinta en distintos ejercicios que un mismo individuo realice. También la capacidad de clasificar imágenes bajo conceptos léxicos parece que dependa del ejercicio en concreto. Si en un ejercicio de clasificación se eligen determinadas imágenes para realizar el aprendizaje supervisado que realiza la máquina, su resultado será distinto al que obtendríamos realizando otro ejercicio con distinta elección de imágenes. En este sentido, las imágenes elegidas para aprehender la capacidad en un ejercicio, determinarán la propia habilidad. ¿Es el que elige las imágenes el po-

seedor del concepto intrínseco a la habilidad? Por supuesto que sí, si no, no sería capaz de proporcionar imágenes adecuadas para el aprendizaje, pero, ¿esto invalida la capacidad de la máquina para aprehender el concepto y mostrar su habilidad de usarlo en la clasificación de imágenes? Entendemos que no. Podemos argumentar en contra que la habilidad de usar el concepto adecuadamente para la clasificación no supone la aprehensión total del concepto, pero sí muestra una aprehensión para la tarea de clasificación.

Si aceptamos que en el experimento los humanos clasificadores poseen el concepto a clasificar, es decir tienen la habilidad de usarlo tanto en el lenguaje como en las diferentes tareas que ejecuten entre las que pueda estar la clasificación de imágenes; si aceptamos que las tasas de aciertos de los humanos en la tarea de clasificación del experimento es propia de un agente del que se sabe que posee el concepto; y si la máquina muestra una tasa de errores similar a la de los humanos -que poseen el concepto- en la tarea de clasificación vista en el experimento, ¿por qué no aceptar que la máquina posee el concepto al menos para la tarea de clasificación del experimento? Desde una posición disposicionalista, definiendo el concepto como una habilidad, parece que la máquina sí posee el concepto, al menos para tareas de clasificación.

Si en humanos, la posesión de un concepto es una función de la habilidad que tienen de usarlo en distintas tareas, el grado de posesión dependerá de la habilidad de su uso en diferentes tareas. ¿Cuántas tareas?, ¿hay un límite mínimo de tareas para determinar la posesión de un concepto?, ¿hay tareas que sean necesarias y suficientes para evaluar la posesión de un concepto? Sí parece necesaria la identificación de la extensión del concepto, y la habilidad de clasificación vista en el experimento lleva implícita cierta habilidad en dicha identificación. Además, si dada una imagen, sería capaz de identificarla apelando a qué conceptos ejemplifica, muestra cierta habilidad de identificar imágenes apelando a los conceptos que instancian. El programa podría hacerse más complejo de modo que la máquina clasifique las imágenes según diversos conceptos (diversos experimentos, en realidad) y luego combine los resultados. Es un tema que escapa a la presente tesis.

Aún nos puede despertar dudas la equivalencia entre conceptos y habilidades. ¿Hay diferencias en el uso cotidiano entre concepto y habilidad? Parece que sí según (Glock 2010).

- *Primero*: Cuando pensamos en un concepto suele ser para definirlo o explicarlo, pero definir o explicar un concepto no es definir o explicar una habilidad, ni siquiera una capacidad. Explicar una habilidad es explicar sus pre-condiciones causales mientras que explicar un concepto es explicar su contenido. Definir una habilidad es definir una 'habilidad de hacer', en este sentido, la habilidad está individualizada a través de su ejercicio. Para explicar un concepto basta con especificar las condiciones que un objeto debe satisfacer para 'caer' dentro del concepto.

- *Segundo*: Un concepto puede ser instanciado o satisfecho por 'entidades' que satisfacen las condiciones para 'caer' dentro del concepto. Esas 'entidades' no pueden entenderse como habilidades, o al menos, no en el mismo sentido.

- *Tercero*: Los conceptos tienen una extensión (los objetos que 'instancian' el concepto) y una intensión (las características que califican a los objetos para 'instanciar' un concepto), y esto no se puede decir de las habilidades. La habilidad que supondría poseer el concepto F tiene una extensión, pero no es el rango de 'cosas que son F'. Por un lado tenemos los sujetos que poseen la habilidad del concepto F y por otro el rango de las situaciones en las cuales estos poseedores pueden aplicar la habilidad del concepto F. No parecen tener ni la misma extensión ni la misma intensión.

- *Cuarto*: Un concepto puede estar en una oración sin apelar a nada externo a sí mismo. Por ejemplo, el concepto 'duro' en la oración 'Juan trabaja duro' solo necesita la palabra 'duro'. Pero una habilidad solo puede estar siendo mencionada la habilidad. Por ejemplo:

  "... la habilidad de mentir convincentemente es una gran cualidad en los negocios ..."

Entonces, ¿podemos afirmar que poseer un concepto es poseer una habilidad?

(i) Poseer un concepto (φ) $\equiv$ Poseer una habilidad (P)

Glock afirma que no podemos establecer el principio general

(ii) Tener (φ) $\equiv$ Tener (P) $\Rightarrow P = \varphi$

a este caso. En este caso, (i) no puede ser desglosado como

(i') S tiene el concepto $\varphi \Longleftrightarrow S$ tiene la habilidad de operar con φ

Entonces, si tener un concepto es una habilidad, esta es una habilidad de operar con el concepto, el concepto por sí mismo no puede ser idéntico a la habilidad. El concepto es empleado algunas veces para ejercitar la habilidad. Si aceptamos que los conceptos son una clase de herramienta cognitiva, los conceptos son 'cosas' empleadas en el ejercicio de habilidades conceptuales, lo mismo que las herramientas son 'cosas' empleadas en el ejercicio de habilidades manuales o técnicas. Aun así hay diferencias entre poseer una herramienta y poseer la habilidad de emplearla. Y esto no es aplicable a los conceptos, ya que poseerlos implica poseer la habilidad de usarlos.

Si consideramos a la herramienta como una técnica y no como un objeto, entonces dominar o poseer una técnica es dominar o poseer una habilidad. Si los conceptos son técnicas para usar palabras y otras operaciones mentales con o sin lenguaje. Entonces ¿qué clase de operaciones mentales? Una contestación plausible es que el pensamiento conceptual gira en torno a la clasificación y la inferencia.

Glock en (Glock 2010) considera que un concepto no es idéntico a la capacidad de clasificar o inferir, sino solo con las técnicas empleadas por alguien que ejerce la habilidad de clasificar o inferir. Y estas técnicas se basan en unas reglas o principios, por lo cual concluye que los conceptos son reglas o principios de clasificación y/o inferencia.

En nuestro experimento las reglas de clasificación están determinadas por la función de distribución, los vectores beta y el límite de equiprobabilidad. Nada podemos decir de las inferencias.

Concluyendo, poseer un concepto φ implica poseer al menos una habilidad particular, a saber, ser capaz de reconocer qué objetos, situaciones, ejemplifican el concepto φ. Esto supone atribuir a quien posee un concepto una capacidad discriminativa. Si x no manifiesta esa capacidad (porque comete demasiados errores a la hora de discernir instanciaciones del concepto φ) no parece que quepa atribuirle el concepto. Por tanto, la capacidad discriminativa es una condición necesaria para la atribución de un concepto. Así, siendo:

- *(1)* "x posee el concepto φ"

- *(2)* "x posee la capacidad P de discernir las instancias de φ de las instancias de no-φ"

cabe admitir que (1) implica lógicamente (2). Ahora bien, ¿qué diríamos de la inversa? ¿Es (1) implicada lógicamente por (2)? ¿Es la capacidad discriminativa suficiente para atribuir el concepto? La respuesta que nosotros vamos a dar a esta pregunta es matizada.

Al referirse a la función desempeñada por los conceptos se alude a menudo a la categorización. Naturalmente que los conceptos son usados para clasificar. Esa clasificación superpone diferentes niveles con sus correspondientes relaciones de inclusión (por ejemplo: perros, gatos, animales domésticos,...), con lo cual se abre la posibilidad de realizar inferencias. Nuestra máquina no realiza inferencias porque el algoritmo no persigue tal cosa. Es dudoso incluso que pueda decirse que la máquina categoriza, salvo que una simple clasificación dicotómica φ / no-φ pueda considerarse una categorización.

Algunos autores han sostenido que la atribución del concepto exige la capacidad de realizar inferencias (por ejemplo, Davidson para *el argumento del 'holismo'* v. supra 3.1, p. 70). Demandarían, pues, que x sea capaz de realizar inferencias en las que emplee φ, y eso nuestra máquina no lo hace. De acuerdo con tal criterio, la máquina no posee el concepto φ. Sin embargo, en el proceso descrito en capítulos anteriores, la máquina aprende algo. Se nos dirá, tal vez, que ha adquirido una capacidad discriminativa concreta, pero que eso no es adquirir un concepto. Nuestra respuesta a esto es que, ciertamente, la máquina posee una capacidad discriminativa que antes no tenía, que se manifiesta en la clasificación y reconocimiento de imágenes. Es importante resaltar que dicha capacidad se adquiere gracias a la codificación de una serie de parámetros (los vectores de descriptores, los centroides, el vector beta, el límite de equiprobabilidad), parámetros que la máquina es capaz de reajustar si se aumenta el input de imágenes. Esto no puede ser obviado, en nuestra opinión, y plantea la cuestión de qué es lo que ha adquirido la máquina, algo que no es simplemente, por lo que acabamos de decir, una capacidad discriminativa.

Desde luego, no queremos defender una concepción de cariz pragmatista que identifica los conceptos con capacidades, sean estas discriminativas, de categorización o inferenciales, o con un cúmulo de disposiciones que el hablante manifiesta en ciertas condiciones. Aparte de las dificultades que conlleva tal perspectiva, no

nos es necesario comprometernos con ella. Y es que, nótese que, ni siquiera aunque se mantenga que

- (3) Tener una capacidad $P \Longleftrightarrow$ Tener un concepto φ,

afirmación mucho más fuerte de lo que nosotros aceptaríamos[32], de ahí no se sigue que

- (4) Capacidad $P \equiv$ concepto φ.

Entonces, si en este caso (4) no se sigue de (3), mucho menos se seguirá de la posición aquí defendida por nosotros, a saber, que "Tener un concepto $\varphi \rightarrow$Tener una capacidad P", pero no consideramos que "Tener una capacidad $P \rightarrow$ Tener un concepto φ".

Recapitulemos. La máquina ha adquirido una habilidad. Dicha habilidad –discriminativa, en particular- se sustenta en una codificación abstracta, pero revisable en función del input de entrada. Por otro lado, atribuir a la máquina la posesión de un concepto en sentido pleno, como lo haríamos, por ejemplo, en el caso de un agente humano que no solo manifiesta su habilidad discriminativa, sino que también realiza inferencias, quizá sea ir demasiado lejos, incluso aunque se trate de un concepto estrechamente vinculado a determinados contenidos perceptivos.

En este punto, para dar cuenta del aprendizaje de la máquina, la disyuntiva que se nos presenta es:

- (a) atribuir la adquisición de algo que no es un concepto;
- (b) atribuir la posesión de un concepto (en particular el concepto 'luz-diurna') no en un sentido pleno, sino parcial.

Los resultados obtenidos en el experimento no apuntan concluyentemente hacia ninguna de las dos opciones, ya que estamos más bien confrontando diferentes interpretaciones de los resultados obtenidos. ¿Por cuál de ambas nos inclinamos nosotros? En nuestra opinión, debe preferirse la opción (b). A continuación trataremos de justificar por qué.

[32]Pues admitimos que "Tener un concepto $\varphi \rightarrow$Tener una capacidad P", pero no consideramos que "Tener una capacidad $P \rightarrow$ Tener un concepto φ".

Lo que la máquina ha aprendido a hacer es un componente esencial de lo que un humano aprende cuando adquiere un concepto, esto es, una habilidad para discriminar, y con ello un conocimiento sobre qué propiedades son relevantes y cuáles no. Es verdad que la máquina no tiene por qué guiarse por las mismas propiedades que guían al sujeto humano. Así, este aprecia que en la imagen aparece un vagón de metro, y por tanto que la luz no es natural, mientras que la máquina no percibe un vagón de metro. Su detección se basa en criterios diferentes, en realidad propiedades ligadas al color, brillo y textura de las imágenes, que no incorporan un contenido semántico. No obstante, los resultados obtenidos por la máquina revelan que esta domina al menos la primera fase en la adquisición de un concepto. Esto no permite hablar de una adquisición plena del concepto, pero el algoritmo discriminativo podría ser implementado con un módulo que permitiera categorizaciones superpuestas sobre los conceptos aprendidos en la fase inicial (por ejemplo, 'franja-del-día', o 'animal doméstico'), y que habilitara para la realización de inferencias[33] (v. supra p. 118). Por eso preferimos hablar aquí de una adquisición parcial del concepto, susceptible de ser gradualmente reforzada en caso de que sea implementada y de que tal implemento arroje resultados satisfactorios, en vez de considerar lo que la máquina ha aprendido nada tiene que ver con el concepto 'luz-de-día'.

Esto significa que la adquisición de un concepto (y consiguientemente, la atribución de un concepto por parte de un sujeto a otro, en este caso, a una máquina) no es un asunto que solo admita un sí o un no por respuesta. Cabe hablar, pues, de una adquisición/atribución parcial, que en un caso como el que nos ocupa nos parece más justificada que el resto de alternativas o interpretaciones disponibles.

[33] Por ejemplo el NeuralTalk2 (https://github.com/karpathy/neuraltalk2 consultado en abril de 2016), que utiliza los algoritmos matemáticos del paquete Torch (http://torch.ch/ consultado en abril de 2016); o el "Google's Brain-Inspired Software (https://www.technologyreview.com/s/532666/googles-brain-inspired-software-describes-what-it-sees-in-complex-images/ consultado en abril de 2016)

Esto no es el fin, ni siquiera es el comienzo del final. Pero, posiblemente, sea el fin del comienzo.

Winston Churchill

CAPÍTULO

5

Conclusiones: ¿una máquina aprendiendo conceptos?

¿Qué tipo de conceptos?

Entendemos como máquina solamente este programa de clasificación de imágenes. En este sentido, limitamos la capacidad visual a la visión bidimensional estática.

Desde el punto de vista de las teorías representacionales, y según la clasificación de grados de "conceptos" comentada en el capítulo 3.1, p.71, por E. Camp (Camp 2009) , debemos conformarnos con el concepto "minimalista", ya que aunque podamos atribuir a la máquina tareas básicas de conceptos como la representación y un razonamiento mínimo sobre datos (la habilidad de clasificación), no podemos atribuirle la independencia del estímulo, es decir, la habilidad de clasificar imágenes bajo conceptos es completamente pasiva, necesita un estímulo externo, la imagen, para poder realizar su trabajo de clasificación.

Con esta visión "minimalista", los elementos candidatos a formar un concepto derivado del experimento son, por un lado, elementos modales como el conjunto de imágenes de entrenamiento, junto con sus vectores de descriptores y su representación en los espacios HSV y/o RGB, los centroides y sus vectores de descriptores,

las características virtuales y los elementos aprendidos para la clasificación (elementos que ya no pueden considerarse como 'no conceptuales'), el vector beta y el límite de equiprobabilidad. Por otro lado, debemos agrupar junto a estos elementos un elemento amodal, el símbolo léxico que representa al concepto. Este será un buen candidato para representar la interfaz de comunicación entre los vehículos de representación modal y amodal.

Estos elementos contribuyen en la tarea de identificación o clasificación de una instancia bajo un concepto léxico, pero no con la recreación visual de un concepto – por el problema de la irreversibilidad -, también faltaría determinar la individuación discursiva dentro de un escenario de pensamiento, es decir, ¿qué imagen de un concepto proyectamos ante un determinado escenario de pensamiento? Para dar respuesta a esta cuestión deberíamos plantearnos la construcción de redes semánticas tanto desde las características virtuales como desde las redes semánticas de conceptos léxicos, que escapan del alcance de esta tesis.

En cuanto a su representación, la teoría del prototipo cubre las expectativas si se elige el centroide de positivas como prototipo, con las matizaciones sobre la falta de significado de sus características comentadas en (capítulo 2.4, p. 56; capítulo 3.2, p. 78). Como ya se indicó, esta falta de significado de las características virtuales, junto con su buen resultado apuntan a reflexionar sobre la utilidad de la variable representacional (su significado). Por otro lado, los elementos aprendidos (vector beta y límite de equiprobabilidad), también pueden ser considerados como la representación del concepto en la teoría del prototipo como se discutió en (v. supra capítulo 3.2, p. 84 y notas 16, 17). Aunque el vector beta no parece que tenga un significado semántico representacional al estilo humano, sí que cumple con el requisito de medir el grado de similitud de una instancia respecto al concepto. Además, este vector junto con el límite de equiprobabilidad determina si una imagen instancia o no un determinado concepto. Tampoco parece que los proxitipos, propuestos por neoempiristas como Prinz, resuelvan el problema de la representación. Por una parte pueden considerarse como una extensión de las características de la teoría del prototipo, añadiendo nuevas variables que en nuestro caso no resuelven la recombinabilidad (composicionalidad) sistemática exigida por Evans ni la reactivación de los circuitos modales de los conceptos en ausencia de entrada perceptiva (v. supra 4.1, p. 104). Por otra parte, también serían necesarias estas

redes semánticas para iniciar procesos de inferencias, para la recombinabilidad y para tareas explicativas de contenido semántico.

En resumen, desde el punto de vista de las teorías representacionales, podemos admitir que estamos ante un modelo de adquisición parcial de conceptos, que es capaz de clasificar pero no puede realizar inferencias ni tiene capacidad explicativa.

Desde el punto de vista de las teorías disposicionalistas, la atribución de conceptos a animales, según la versión de Allen discutida, pone en duda la atribución de conceptos cuando solo se dispone de un sentido (como es nuestro caso) Pero, si aceptamos las matizaciones expuestas en (v. supra 4.2, p. 122 y 4.2, p. 123), podemos atribuir a un sujeto O (la máquina) 'razonablemente' la posesión de un concepto φ (timeofday_day) porque cumple con las condiciones exigidas por Allen, con las matizaciones allí expuestas.

También, y siguiendo la discusión realizada en (v. supra 4.2, p. 121), podemos admitir que la habilidad de clasificación de imágenes bajo conceptos léxicos realizada por la máquina evidencia una adquisición al menos parcial del concepto.

Si se ha de optar por una de las dos alternativas, representacional o disposicional, después de las discusiones expuestas en esta tesis, entendemos que la alternativa disposicional cuadraría mejor con los parámetros del experimento, entre otras razones, por la falta de explicaciones semánticas de los modelos representacionales expuestos.

¿Un ordenador/una máquina aprendiendo 'conceptos'?

Y bien, ¿podemos afirmar entonces que, cuando el experimento arroja una tasa de aciertos predictivos razonablemente alta, el ordenador *ha aprendido* el concepto que subyace al proceso de etiquetaje humano de la base de datos? ¿El concepto que se aprende es, estrictamente, una entidad matemática que define una imagen virtual? ¿O más bien, a juzgar por el éxito predictivo incluso superior al del agente humano, lo que se ha aprehendido, a través de imágenes, es la interpretación "humana" del concepto?.

Se puede objetar que el programa falla demasiado como para considerar que ha aprendido el concepto. Pero hay que tener en cuenta, primero, que los vínculos concepto-imagen planteados en el experimento son bastante complejos, como ya dijimos en su momento. No se trata de 'perro' o 'gato', sino de conceptos como 'timeofday_day' que pueden ser instanciados o no dependiendo de sutilezas de la

imagen (tonalidades de color, por ejemplo) y que pueden dar lugar a interpretaciones diferentes sobre si una imagen contiene o no el concepto. Esta complejidad también han de afrontarla los sujetos humanos (ya que no interviene solo uno) que han de realizar la clasificación modelo, la que se admite que es portadora de verdad en el experimento. Por eso se debería comprobar si los sujetos humanos han aprendido el concepto para poder realizar la clasificación objetivamente. La ignorancia y el error, sobre todo para los casos atípicos, afligen a los humanos en estas tareas. Los errores cometidos en la fase de clasificación humana pasarán a construir el etiquetaje de referencia, supuestamente portador de verdad, en el experimento. Pero si la máquina toma como modelos imágenes erróneas, esos errores se trasladarán al *output*, incrementando la tasa de fallos. Además, el experimento no trata de simular una clasificación como lo haría un humano, sino desarrollar un método alternativo y comparar estadísticamente los resultados. En este sentido no cabe comparar si la máquina y el humano fallan en las mismas instancias, ya que sus métodos de clasificación son diferentes, sino comparar su eficiencia en la tarea de clasificación.

Podemos admitir, pues, que el programa ha aprendido a clasificar instancias bajo conceptos que presentan una complejidad doble. Por un lado la que se deriva del hecho de que pequeñas diferencias en la imagen generan diferencias conceptuales relevantes; y por otro, la complejidad interpretativa humana, en parte consecuencia de lo anterior, pero no solo de eso. El hecho es que el programa responde con una tasa de aciertos al menos similar a la que obtendría un humano –aunque no disponemos de todos los datos empíricos del proceso de clasificación realizado por los humanos, sí sabemos que en cada concepto ha intervenido más de un humano, y que la divergencia entre ellos ha sido alta, con lo que se optó por considerar como imágenes positivas, todas aquellas que algún agente humano hubiera clasificado como tales.

Entonces, si admitimos que el programa es capaz de clasificar instancias -imágenes- bajo un determinado concepto, entendiendo este concepto como un complejo interpretativo multihumano de un concepto léxico, ¿se podría inferir de esto que el ordenador ha aprendido el concepto 'timeofday_day', por ejemplo?.

Si consideramos los conceptos como habilidades, parece que el ordenador ***sí tiene la habilidad de clasificar instancias -imágenes- bajo conceptos***. ¿Qué conceptos? Los que subyacen a la clasificación de imágenes que aportan los humanos

participantes en el experimento.

Pero esto, podría argüirse, no es lo mismo que aprender el concepto. Clasificar imágenes es uno de los ingredientes, un elemento necesario tal vez – y es que seguramente de quien cometa bastantes fallos a la hora de distinguir entre imágenes de perros y de gatos, no diríamos que ha adquirido el concepto 'perro'–, pero no suficiente. Ciertamente, una habilidad que no se ha considerado en esta tesis es la de realizar *inferencias*. Quien entiende el concepto 'perro' sabe clasificar ciertas imágenes, pero también sabe que un perro es un animal, y de ahí que pueda realizar ciertas inferencias ("Los perros son mortales," por ejemplo).

Para probar la capacidad de realizar inferencias debemos plantearnos el procesamiento simbólico a partir de los conceptos aprendidos como habilidades de clasificación, de esta forma, si unimos todos los conceptos que "caen" bajo una familia, por ejemplo "timeofday", podemos inferir el concepto "timeofday" como la unión de sus conceptos constituyentes, "day", "night" y "sunrisesunset". En este sentido, el aprendizaje de "timeofday" se habría realizado en un plano de abstracción diferente a la mera clasificación, y estaríamos hablando de procesamiento simbólico a partir de conceptos aprendidos sensorialmente. En inteligencia artificial existen modelos que soportan bien este paradigma y que se han popularizado en los últimos años, sistemas expertos y sistemas basados en conocimiento. Este tema no ha sido tratado aquí, ya que el objetivo de esta tesis consiste en abordar la habilidad clasificadora y su relación con el aprendizaje conceptual, interpretar a partir de estas evidencias qué tipo de conceptos puede tener esta máquina, y finalmente ver su encaje en las teorías de conceptos dentro de la filosofía y las ciencias cognitivas. Mientras que el humano, cuando aprende a clasificar puede estar aprendiendo también a realizar inferencias, en este experimento solamente se evidencia la capacidad de clasificar imágenes. El ordenador evidencia una alta *capacidad discriminativa*, y en esta tarea llega a aventajar al humano. Ahora bien, si la adquisición de un concepto 'visual' requiere otras capacidades –como *la inferencial*–, nuestro experimento nada puede decir sobre ello.

En esta tesitura podríamos tal vez hablar de distintos grados en el aprendizaje/adquisición de un concepto. La capacidad de discriminación sería un primer nivel mientras que la capacidad inferencial constituiría un nivel superior. Y aún podría señalar alguien la vinculación con la acción, como una fase adicio-

nal, exigiendo para el dominio pleno del concepto la inserción del sujeto en una práctica social. En todo caso, el proceso de aprendizaje –en cuanto capacidad discriminativa- de nuestro ordenador, basado en regresiones logísticas a partir de ejemplos suministrados por humanos, es muy distinto a los modelos comúnmente empleados en filosofía y psicología para describir cómo operan los humanos (e incluso los animales). A pesar de todo pensamos que los resultados empíricos obtenidos invitan a explorar seriamente esta alternativa, una alternativa que queda ubicada en las coordenadas de la teoría del prototipo, en cuanto a la estructura de los conceptos, y la identificación del concepto con una capacidad/habilidad, una disposición desde un punto de vista ontológico.

¿Es necesaria otra teoría?

Aunque los sujetos que pueden poseer conceptos, según las ciencias cognitivas y la filosofía son los humanos adultos, y con algunas matizaciones, niños y animales, algunas de las características exigidas por las teorías sobre conceptos también las cumple nuestro programa. En este sentido, no necesitamos más teorías para admitir que nuestro experimento pone de manifiesto una aprehensión parcial de un concepto,- concepto en el sentido que lo tendría un humano adulto-, y que pueden rediseñarse nuevos experimentos con procesamiento simbólico en los cuales poder demostrar la aprehensión de conceptos desde dos vías, una sensorial, como la expuesta en la presente tesis, y otra por mecanismos amodales simbólicos, tipo lenguaje, que construyan redes semánticas.

Cuanto mayor es la dificultad, mayor es la gloria.

Marco Tulio Cicerón

CAPÍTULO

6

Anexos

6.1 *Anexo I: Aparato matemático de la teoría.*

6.1.1 *El modelo logit*

El modelo predictivo que subyace bajo las regresiones logísticas trata de ajustar una serie de parámetros para conseguir que clasifique adecuadamente un conjunto de instancias positivas y negativas proporcionadas para construir la función de probabilidad:

$$p_i = F(\beta_0 + \overrightarrow{\beta_i} * \overrightarrow{X_i})$$

Donde F ha de pertenecer a la familia de funciones no decrecientes acotadas entre cero y uno, es decir, una función de distribución. Si tomamos F como la función de distribución logística dada por:

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \overrightarrow{\beta_i} * \overrightarrow{X_i})}}$$

Esta función tiene la ventaja de ser continua y además encontramos que:

$$1 - p_i = \frac{e^{-(\beta_0 + \overrightarrow{\beta_i} * \overrightarrow{X_i})}}{1 + e^{-(\beta_0 + \overrightarrow{\beta_i} * \overrightarrow{X_i})}} = \frac{1}{1 + e^{+(\beta_0 + \overrightarrow{\beta_i} * \overrightarrow{X_i})}}$$

Si planteamos la transformación g como el logaritmo del cociente de las probabilidades de los sucesos, podemos obtener un modelo lineal más tratable. El hecho de maximizar la probabilidad "p" de que una imagen "i" sea positiva -que se acerque a 1- es lo mismo que maximizar la función g así definida -que se acerque a infinito-:

$$g_i = \log\left(\frac{p_i}{1 - p_i}\right) = \log\left(\frac{\frac{1}{1+e^{-(\beta_0 + \overrightarrow{\beta_i} * \overrightarrow{X_i})}}}{\frac{e^{-(\beta_0 + \overrightarrow{\beta_i} * \overrightarrow{X_i})}}{1+e^{-(\beta_0 + \overrightarrow{\beta_i} * \overrightarrow{X_i})}}}\right) = \log\left(\frac{1}{e^{-(\beta_0 + \overrightarrow{\beta_i} * \overrightarrow{X_i})}}\right) = \beta_0 + \overrightarrow{\beta_i} * \overrightarrow{X_i}$$

La variable g representa así la diferencia de probabilidades de pertenecer a positivas o negativas en escala logarítmica de la imagen "i". El hecho de ser lineal a las variables explicativas beta facilita la estimación e interpretación del modelo.

Por otro lado, la obtención de los parámetros beta, o vector beta, se consigue al maximizar la función:

$$\prod_{i=1}^{k} g_i * \prod_{j=1}^{J} g_j \text{ donde } i \in positivas \wedge g_i = \log\left(\frac{p_i}{1 - p_i}\right) \wedge j \in negativas \wedge g_j = \log\left(\frac{1 - p_j}{p_j}\right)$$

Una vez estimado el modelo, esto es, conocidos los parámetros beta, ya somos capaces de construir un predictor con ellos que, a partir de las características nos diga la probabilidad que tiene una instancia imagen de contener el concepto objeto de la función, sin más que aplicar la fórmula F.

6.1.2 *Transformación del espacio RGB a HSV*

La función de transformación de RGB a HSV es la siguiente:

Sea MAX el valor máximo de los componentes (R, G, B), y MIN el valor mínimo de esos mismos valores, los componentes del espacio HSV se pueden calcular como muestra la tabla 6.1.

Canal	Función de transformación	Condiciones de aplicación
H=	No definido	Si MAX=MIN
	$60° * \frac{(G-B)}{(MAX-MIN)} + 0°$	Si MAX=R y G $\geq$ B
	$60° * \frac{(G-B)}{(MAX-MIN)} + 360\circ$	Si MAX=R y G ¡B
	$60° * \frac{(B-R)}{(MAX-MIN)} + 120°$	Si MAX=G
	$60° * \frac{(R-G)}{(MAX-MIN)} + 240°$	Si MAX=B
S=	0	si MAX=0
	$1 - \frac{MIN}{MAX}$	en otro caso
V=	MAX	Siempre

Cuadro 6.1: Transformación de espacio RGB a HSV.

6.2 *Anexo II: Explicaciones y predicciones del modelo.*

El formato de presentación de los resultados se desglosa en una tabla por cada concepto. En cada tabla, dos supercolumnas indican los valores de las explicaciones y los valores de las predicciones. En cada supercolumna, encontramos las columnas que nos indican, la baseline, la equiprobabilidad, las positivas y las negativas en términos absolutos.

Por cada bloque de equiprobabilidad-Baseline, encontramos filas distintas que nos indican los aciertos y fallos en la clasificación de las imágenes, y también los porcentajes o tasas de acierto, tanto para imágenes positivas, negativas, como las globales.

Solamente se muestra el concepto 0 en las tablas 6.2 y 6.3, el resto de conceptos en ficheros adjuntos.

Concepto 0: timeofday_day

Concepto 0: timeofday_day . Baseline 1							
		Explica			Predice		
Bsl[34]	Equipr[35]		Pos[36]	Neg[37]		Pos[38]	Neg[39]
1	0,5	Bien	10	60	Bien	394	9570
		Mal	10	0	Mal	4503	533
		Explica %	50	100	Predice %	8,05	94,72
		Explica Tot %	**87,5**		Predice Tot %	**66,43**	
	0,3188 0995	Bien	18	59	Bien	2192	6456
		Mal	2	1	Mal	2705	3647
		Explica %	90	98,34	Predice %	44,76	63,90
		Explica Tot %	**96,25**		Predice Tot %	**57,65**	

Cuadro 6.2: Resumen estadístico de explicaciones y predicciones para el concepto 0. Mejor modelo explicativo Baseline 1, primera aproximación (96,25 %). Mejor modelo predictivo Baseline 1, tercera aproximación (66,42 %).

[34]Baseline
[35]Equiprobabilidad
[36]Del conjunto de Positivas
[37]Del conjunto de Negativas

Concepto 0: timeofday_day . Baseline 2 y 3							
		Explica			Predice		
Bsl[40]	Equipr[41]		Pos[42]	Neg[43]		Pos[44]	Neg[45]
2	0,5	Bien	0	60	Bien	1456	7301
		Mal	20	0	Mal	3441	2802
		Explica %	0	100	Predice %	29,73	72,27
		Explica Tot %	**75**		Predice Tot %	**58,38**	
	0.1478 7856	Bien	11	59	Bien	4686	454
		Mal	9	1	Mal	211	9649
		Explica %	55	98,34	Predice %	95,69	4,49
		Explica Tot %	**87,5**		Predice Tot %	**34,27**	
3	0,5	Bien	0	60	Bien	567	9165
		Mal	20	0	Mal	4330	938
		Explica %	0	100	Predice %	11,58	90,72
		Explica Tot %	**75**		Predice Tot %	**64,88**	
	0.1493 93585	Bien	12	48	Bien	4079	1888
		Mal	8	12	Mal	818	8215
		Explica %	60	80	Predice %	83,3	18,69
		Explica Tot %	**75**		Predice Tot %	**39,78**	

Cuadro 6.3: Resumen estadístico de explicaciones y predicciones para el concepto 0. Mejor modelo explicativo Baseline 1, primera aproximación (96,25 %). Mejor modelo predictivo Baseline 1, tercera aproximación (66,42 %).(bis)

[38] idem 1
[39] idem 2
[40] Baseline
[41] Equiprobabilidad
[42] Del conjunto de Positivas
[43] Del conjunto de Negativas
[44] idem 1
[45] idem 2

6.3 *Anexo III: Valores de corte de equiprobalidad para aproximación 1.*

Los valores de corte para la aproximación primera se muestran en la tabla 6.4.

Concepto	Baseline 1	Baseline 2	Baseline 3
0	0.31880995	0.14787856	0.149393585
1	0.32473215	0.171522	0.22171294
2	0.3432363	0.137789775	0.1506634
70	0.3198159	0.170178195	0.10495042
71	0.2905345	0.150273855	0.163734325
72	0.3215002	0.12114043	0.197789615
73	0.2971487	0.1578047	0.159277215

Cuadro 6.4: Valores para la equiprobabilidad según aproximación primera.

6.4 *Anexo IV: Semántica de los conceptos del experimento*

La organización Imagecleff suministró a los clasificadores humanos la siguiente semántica de conceptos que han sido tratados en la tesis.

0 timeofday_day

The picture shows that it was taken during the day.

1 timeofday_night

The picture shows that it was taken during the night.

2 timeofday_sunrisesunset

The picture shows that it was taken during the transition from night to day or from day to night, i.e. during sunrise, sunset, dusk, dawn or twilight.

70 view_portrait

The picture shows a scene where one or more persons are the center of attention, typically facing the camera and aware that a photo is being taken of them. The photo normally captures at least their entire face, although a small part of it may be missing.

71 view_closeupmacro

The picture shows a close-up of objects, where a lot of zoom has been used by the photographer, and includes macro shots, where things are shown much larger than they normally are. In contrast with a portrait a close-up can be of anything and not just people, although a photo showing only a face would be called a portrait.

72 view_indoor

The picture shows an indoor scene.

73 view_outdoor

The picture shows an outdoor scene.

6.5 *Anexo V: Estadísticas de los agentes humanos*

La organización Imagecleff facilita la siguiente información sobre los clasificadores humanos de conceptos. Dice textualmente en (http://www.imageclef.org/2012/photo-flickr/dataset):

> **Concept features**
>
> We have solicited the help of workers on the Amazon Mechanical Turk platform to perform the concept annotation for us. To ensure a high standard of annotation we used the CrowdFlower platform that acts as a quality control layer by removing the judgments of workers that fail to annotate properly. We reused several concepts of last year's task and for most of these we annotated the remaining photos of the MIRFLICKR-25K collection that had not yet been used before in the previous task; for some concepts we reannotated all 25,000 images to boost their quality. For the new concepts we naturally had to annotate all of the images.
>
> - ***Concepts*** For each concept we indicate in which images it is present. The 'raw'concepts contain the judgments of all annotators for each image, where a '1'means an annotator indicated the concept was present whereas a '0'means the concept was not present, while the 'clean'concepts only contain the images for which the majority of annotators indicated the concept was present. Some images in the raw data for which we reused last year's annotations only have one judgment for a concept, whereas the other images have between three and five judgments; the single judgment does not mean only one annotator looked at it, as it is the result of a majority vote amongst last year's annotators.
> - ***Annotations*** For each image we indicate which concepts are present, so this is the reverse version of the data above. The 'rawánnotations contain the average agreement of the annotators on the presence of each concept, while the 'cleanánnotations only inclu-

> de those for which there was a majority agreement amongst the annotators.
>
> You will notice that the annotations are not perfect. Especially when the concepts are more subjective or abstract, the annotators tend to disagree more with each other. The raw versions of the concept annotations should help you get an understanding of the exact judgments given by the annotators.

Por tanto, hay clasificadores que han sido aparados por el sistema por no ajustarse a los criterios de convergencia mínimos exigidos por la plataforma informática. Además, muchos conceptos han sido etiquetados solamente por un clasificador, con lo que nadie garantiza que se ajusten a los criterios de convergencia en este caso, al no haber competencia en la clasificación. Aun así, la información que ha servido de base para esta competición se puede resumir como se muestra en las tablas (v. infra 6.5, 6.6, 6.7, 6.8, 6.9, 6.10, 6.11).

Podemos apreciar que más de un 70 % de las imágenes para cada concepto han sido etiquetadas sólo por un miembro humano, sujeto a la sospecha de no haber podido ser evaluada su eficiencia por el sistema. Además, para las imágenes que solo cuentan con una opinión, esta no necesariamente es del mismo clasificador, con lo que el problema de su calidad se vuelve más inestable.

Por otra parte, de las imágenes que cuentan con más de una opinión, sabemos que han sido excluidos los individuos que por motivos de divergencia no alcanzan el mínimo exigido por el control de calidad del sistema informático mencionado.

Así, solamente un máximo del 30 % de opiniones pueden contrastarse entre los individuos humanos clasificadores, y aún así, encontramos divergencias dignas de ser mencionadas, sobre todo cuando hay más de dos opiniones sobre los juicios de clasificación de una imagen. Las gráficas muestran que los consensos arrancan desde el 50 % en el caso del concepto 71 con un porcentaje sobre el global del 28,5 %, hasta el mejor posicionado con un 83,3 en el caso del concepto 2, con un porcentaje sobre el global del 28,9 %. ***Para el concepto 0, los valores son de 59,24 % de consenso, sobre el 29,8 % del total de las imágenes***.

Concepto 0: ”timeofday_day”					
N opiniones	Consenso	Discrepancia	Total	Consenso (%)	Total (%)
1 opinión	10525	0	10525	100	**70,17**
2 opiniones	26	3	29	89,66	0,19
3 opiniones	2426	1700	4126	58,8	27,51
4 opiniones	90	72	162	55,56	1,08
5 opiniones	109	49	158	68,99	1,05
Totales +1op	2651	1824	4475	**59,24**	**29,83**

Cuadro 6.5: Estadísticas de los agentes humanos para el concepto 0 timeofday_day

Concepto 1: ”timeofday_night”					
N opiniones	Consenso	Discrepancia	Total	Consenso (%)	Total (%)
1 opinión	10524	0	10524	100	**70,16**
2 opiniones	25	5	30	83,33	0,2
3 opiniones	3464	655	4119	84,1	27,46
4 opiniones	122	46	168	72,62	1,12
5 opiniones	121	38	159	76,10	1,06
Totales +1op	3732	744	4476	**83,38**	**29,84**

Cuadro 6.6: Estadísticas de los agentes humanos para el concepto 1 timeofday_night

Concepto 2: ”timeofday_sunrisesunset”					
N opiniones	Consenso	Discrepancia	Total	Consenso (%)	Total (%)
1 opinión	10525	0	10525	100	**70,17**
2 opiniones	27	2	29	93,10	0,19
3 opiniones	3655	475	4130	88,50	27,53
4 opiniones	124	34	158	78,48	1,05
5 opiniones	123	35	158	77,85	1,05
Totales +1op	3929	546	4475	**87,80**	**29,83**

Cuadro 6.7: Estadísticas de los agentes humanos para el concepto 2 timeofday_sunrisesunset

Concepto 70: ”view_portrait”					
N opiniones	Consenso	Discrepancia	Total	Consenso (%)	Total (%)
1 opinión	10722	0	10722	100	**71,48**
2 opiniones	1	0	1	100	0,007
3 opiniones	2940	1149	4089	71,90	27,26
4 opiniones	34	35	69	49,28	0,46
5 opiniones	80	39	119	67,23	0,79
Totales +1op	3055	1223	4278	**71,41**	**28,52**

Cuadro 6.8: Estadísticas de los agentes humanos para el concepto 70 view_portrait

Concepto 71: "view_closeupmacro"					
N opiniones	Consenso	Discrepancia	Total	Consenso (%)	Total (%)
1 opinión	10722	0	10722	100	**71,48**
2 opiniones	1	0	1	100	0,007
3 opiniones	2072	2017	4089	50,67	27,26
4 opiniones	24	45	69	34,78	0,46
5 opiniones	59	60	119	49,58	0,79
Totales +1op	2156	2122	4278	**50,40**	**28,52**

Cuadro 6.9: Estadísticas de los agentes humanos para el concepto 71 view_closeupmacro

Concepto 72: "view_indoor"					
N opiniones	Consenso	Discrepancia	Total	Consenso (%)	Total (%)
1 opinión	10734	0	10734	100	**71,56**
2 opiniones	0	0	0	0	0
3 opiniones	2730	865	3595	75,94	23,97
4 opiniones	355	151	506	70,16	3,37
5 opiniones	115	50	165	69,70	1,1
Totales +1op	3200	1066	4266	**75,01**	**28,44**

Cuadro 6.10: Estadísticas de los agentes humanos para el concepto 72 view_indoor

Concepto 73: "view_outdoor"					
N opiniones	Consenso	Discrepancia	Total	Consenso (%)	Total (%)
1 opinión	10734	0	10734	100	**71,56**
2 opiniones	0	0	0	0	0
3 opiniones	2706	889	3595	75,27	**23,97**
4 opiniones	334	172	506	66,01	3,37
5 opiniones	111	54	165	67,27	1,1
Totales +1op	3151	1115	4266	**73,86**	**28,44**

Cuadro 6.11: Estadísticas de los agentes humanos para el concepto 73 view_outdoor

6.6 *Anexo VI: Vectores beta del concepto timeofday_day para Baseline-1*

Concepto 0: 'timeofday_day' en Baseline 1		
Característica	Número de Componentes	vector
1	17, de β_0 a β_{16}	(0,8031207; -19,46222; 5,742456; -2000741; -2,407278; -11,63152; -571534,5; 28,41483; 453,7484; -4402753; -41,49006; 24,675; -1369975; 31,8898; -5,615591; 2970886; -24,95301)
2	15, de β_0 a β_{14}	(3,040603; 11,1785; -2115605; -9,145822; -1452786; 0; -1,56047; 24,32181; -1397408; -0,8246746; -27,57662; 722,6867; -21,80084; 30,09077; -82,1988)
3	17, de β_0 a β_{16}	(0,7862485; -3,982471; -43,48711; 144,5169; -73,37383; -25,01821; 77,98713; -82,93503; 234,6451; 81,76431; -158,8152; 15,08017; -57,65971; -53,63115; -237,8534; 460,6932; -433,2298)
4	17, de β_0 a β_{16}	(-1,141908; -53,02197; 121,8854; -95,49143; -973,2452; 114,25; 8,031661; -36,98705; 43,70482; -1557490; -1960422; 4210265; -27835,22; 757673; -403,9167; -58269,75; -52334,63)
5	17, de β_0 a β_{16}	(-1,011508; -27975,33; 3,457121; -185,4332; -15,35707; -2420501 0; 0; -1540766; 249700,3; -3564016; 0; 0; 1744285; -4914111; -3148208; 3826901)
6	17, de β_0 a β_{16}	(2,70631; -17,69586; -17,58471; -29,56178; 16,36326; -74,40491; 59,51468; -86,63924; 106,7163; 16,03199; -21,74906; -52,49665; -15,05663; -65,72069; 22,07223; 240,4783; -537,8814)

Cuadro 6.12: Vectores del 1 al 6 de 19 para las regresiones del concepto 0, Baseline 1. (1)

Concepto 0: 'timeofday_day' en Baseline 1 (cont. I)		
Característica	Número de Componentes	vector
7	17, de β_0 a β_{16}	(-1,590428; 4,578448; 125,4657; 710,8776; 78,72432; -282,2301; -35,1512; -46,02352; 117,0968; -2430611; 47659,6; 750828,7; -196950,9; 286,9537; 136,7987; -63082,15; 4243888)
8	17, de β_0 a β_{16}	(-1,130179; -2672502; 107,0047; -156,8373; 205,8572; 32760326; 0; 0; -102826434; 5764160; -4269960; 0; 0; 0; -3033514; -271243,1; -3291817)
9	17, de β_0 a β_{16}	(0,1274646; -67,32353; 83,0996; 8,771547; -76,94982; -21,16231; 44,46421; -46,79903; -143,8315; 2,919129; 52,64172; 17,51449; -23,23813; 186,4292; -329,1829; 330,4004; -116,8119)
10	17, de β_0 a β_{16}	(-2,541648; 67,21733; 162,3874; 233,4931; 844,8491; -1203104; 84,01794; -65,49467; 203,6609; -29899,31; -47555,62; 1000357; -14453,26; 5992495; 2090224; -24071,79; 163095,9)
11	17, de β_0 a β_{16}	(-1,01121; -3196113; 27,72675; -728,6156; 4590327; -1252362; -4616822; 0; -48738367; 4842207; 0; 0; 0; 0; -182116,2; -19739,36; -8701113)
12	17, de β_0 a β_{16}	(1,136816; -43,98482; 60,72179; -28,33966; -50,11766; -15,18606; 62,63398; -79,04418; -216,4641; -78,16964; 77,77079 ; -14,0442; -20,04135; 250,9108; -285,5197; 291,3264; -145,5483)

Cuadro 6.13: Vectores del 7 al 12 de 19 para las regresiones del concepto 0, Baseline 1.(2)

Concepto 0: 'timeofday_day' en Baseline 1 (cont. II)		
Característica	Número de Componentes	vector
13	17, de β_0 a β_{16}	(-3,395541; 152,1462; -15,14678; 861,4947; -3291742; -240,0954; 131,3053; -20,32564; 163,4162; -4795819; 205693,2; -51521,59; -2380044; -1251946; 277,5036; -422,3207; -19899175)
14	17, de β_0 a β_{16}	(-1,356175; -47462,73; 195,424; -98,80807; -112,2011; -28703089; 0; 0; 43810494; -3225927; -507322,9; 0; 0; 288653,2 ; -16246,74; 789,0513; 373261,8)
15	17, de β_0 a β_{16}	(-10,43561; -1042855; -50917,54; 4,103413; -4,002319; -84600,74; 215963; -57945,77; 15747,26; -4714693; 2116888; -3257389; 10059,48; -37653,19; 77313,16; 89202,85; -481310)
16	16, de β_0 a β_{15}	(43,49922; -924,7904; 417903; -504,6932; 22,41493; 254,4621; -614177; 413,6458; -621,2029; -6,187164; 7,552834; 121221,3; 97400,01; -568260,9; 150611; 199028,5)
17	17, de β_0 a β_{16}	(-38,89359; -379346,5; -38898,99; 21,58098; -10,7364; 1723649; -1339006; 358306,4; -95915,63; 25682,42; -5718889; -1168595; 4782839; -13842,73; 27936,16; -1833482; 2181765)
18	16, de β_0 a β_{15}	(37,16128; -442,2186; 785,6805; 46,4849; -63,72099; -575624; -864,7836; 1105277; -1921337; 308,9275; -7,627757; -399474,1; -358786,4; 1740302; 127602,7; -1109645)
19	10, de β_0 a β_9	(-2,449304; 371,3339; -196,5908; -134,8638; 199,3459; -64,33069; -16,30434; 30,2763; 8,806394; -26,91127)

Cuadro 6.14: Vectores del 13 al 19 de 19 para las regresiones del concepto 0, Baseline 1.(3)

Los valores del vector beta para cada característica del concepto 0: 'timeofday_day' en Baseline 1 se muestran en las tablas (v. supra 6.12, 6.13 y 6.14).

Bibliografía

[1] Allen, C. (1999). Animal concepts revisited: The use of self-monitoring as an empirical approach. *Erkenntnis*, *51(1)*, 33–40. 122

[2] Allen, C., & Hauser, M. (1996). *Concept Attribution in Nonhuman Animals: Theoretical and Methodological Problems in Ascribing Complex Mental Processes*. Cambridge MA: MIT Press. 121

[3] Bacon, F. (1905[1620]). *The Philosophical Works of Francis Bacon*. London: Routledge.

[4] Barsalou, L. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*, 577–660. 104, 105, 111, 113

[5] Barsalou, L., Simmons, W., Barbey, A., & Wilson, C. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in Cognitive Science*, *7, Issue 2*, 84–91. 109, 113

[6] Beck, J. (2012). The generality constraint and the structure of thought. *Mind*, *124*, 563 – 600. 71, 73

[7] Beckers, T., Miller, R., Houwer, J. D., & Urushihara, K. (2006). Reasoning rats: Forward blocking in pavlovian animal conditioning is sensitive to constraints of causal inference. *Journal of Experimental Psychology: General*, *135(1)*, 1234–1236.

[8] Benavent, J., Benavent, X., de Ves, E., Granados, R., & García-Serrano, A. (2010). Experiences at imageclef 2010 using cbir and tbir mixing information

approaches. *CLEF 2010 LABs and Workshops, Notebook Papers*. 15, 23, 84, 99, 120

[9] Benavent, J., Castellanos, A., Benavent, X., de Ves, E., & García-Serrano, A. (2012). Visual concept features and textual expansion in a multimodal system for concept annotation and retrieval with flickr photos at imageclef2012. *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes*. 15, 23, 84, 120

[10] Benavent, X., García-Serrano, A., Granados, R., Benavent, J., & de Ves, E. (2013). Multimedia information retrieval based on late semantic fusion approaches: Experiments on a wikipedia image collection. *IEEE Transactions on Multimedia*, *15(8)*(1), 2009–2021. 15, 23, 84, 120

[11] Bergman, T., Beehner, J., Cheney, D., & Seyfarth, R. (2003). Hierarchical classification by rank and kinship in baboons. *Science*, *302*, 1234–1236. 69

[12] Bermúdez, J., & Cahen, A. (2015). Nonconceptual mental content. *The Stanford Encyclopedia of Philosophy (Fall 2015 Edition)*. 64, 68

[13] Call, J. (2006). Inferences by exclusion in the great apes: The effect of age and species. *Animal Cognition*, *9*, 393–403.

[14] Camp, E. (2009). Putting thoughts to work: Concepts, systematicity, and stimulus-independence. *Philosophy and Phenomenological Research*, *LXXVIII No. 2*. 71, 74, 133

[15] Castellanos, A., Benavent, J., Benavent, X., García-Serrano, A., & de Ves, E. (2012). Using visual concept features in a multimodal retrieval system for the medical collection at imageclef2012. *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes*. 15, 23, 84

[16] Castellanos, A., Benavent, X., Benavent, J., & García-Serrano, A. (2011). Uned-uv at medical retrieval task of imageclef 2011. *CLEF 2011 Labs and Workshop, Notebook Papers*. 15, 23, 84

[17] Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge: MA: MIT Press. 97

[18] Clark, A. (2013). Expecting the world: Perception, prediction, and the origins of human knowledge. *Journal of Philosophy*, *CX, no. 9*, 469–496.

[19] Collins, A., & Quillian, R. (1972). How to make a language user. *Organization of memory*. 89

[20] Damasio, A. (1989). Time-locked multi-regional retroactivation: A systems-level proposal for the neural substrates of recall and recognition. *Cognition*, *33*, 25–62. 105

[21] Damasio, A., & Damasio, H. (1994). *Cortical systems for retrieval of concrete knowledge: The convergence zone framework*. Cambridge MA: MIT Press. 105

[22] Davidson, D. (1975). Thought and talk. In *Inquiries into Truth and Interpretation*, chap. 11, (pp. 155–170). Oxford: Clarendon Press, second ed. 69, 70

[23] Davidson, D. (2004). What thought requires. In *Problems of rationality*, chap. 9, (pp. 135–149). Oxford: Clarendon Press. 70

[24] de Ves, E., Benavent, X., , Coma, I., & Ayala, G. (2016). A novel dynamic multi-model relevance feedback procedure for content-based image retrieval. *Neurocomputing*, *208*, 99–107. 15, 84

[25] Dennett, D. (1987). *The Intentional Stance*. Cambridge: MA: MIT Press. 90

[26] Deselaers, T., Keysers, D., & Ney, H. (2008). Features for image retrieval: An experimental comparison. *Information Retrieval*, *11*, 77–107. 9

[27] Dove, G. (2009). Beyond perceptual symbols: A call for representational pluralism. *Cognition*, *110*, 412–431. 105, 115, 119, 121

[28] Evans, G. (1982). *The Varieties of Reference*. Oxford: Oxford University Press. 64, 70, 71

[29] Fodor, J. (1984[1975]). *El lenguaje del pensamiento*. Madrid: Alianza. 97, 119

[30] Fodor, J. (1998). *Concepts: where Cogntitive Science Went Wrong*. Oxford: Oxford University Press. 62

[31] Fodor, J. (2007). *Contemporary Debates in Philosophy of Mind*. Oxford: Blackwell. 66

[32] Fodor, J. (2008). *Lot 2: The Language of Thought Revised*. Oxford: Oxford University Press. 66, 97

[33] Gillies, D. (1996). *Artificial Intelligence and Scientific Method*. New York: Oxford University Press. 15

[34] Glock, H. (2010). Concepts, abilities and propositions. *Grazer Philosophische Studien*, *81(1)*, 115–134. 126, 128

[35] Granados, R., Benavent, J., Benavent, X., de Ves, E., & García-Serrano, A. (2011). Multimodal information approaches for the wikipedia collection at imageclef 2011. *CLEF 2011 Labs and Workshop, Notebook Papers*. 15, 23, 84

[36] Griffiths, P. (2009). The distinction between innate and acquired characteristics. *The Stanford Encyclopedia of Philosophy*. 96

[37] Herrnstein, R. (1979). Acquisition, generalization, and discrimination reversal of a natural concept. *Journal of Experimental Psychology: Animal Behavior Processes*, *5*, 116–129. 121

[38] Kenny, A. (2010). Concepts, brains, and behaviour. *Grazer Philosophische Studien*, *81(1)*, 105–113. 125

[39] Kripke, S. (1972). *Naming and necessity*. Harvard University Press. 82

[40] Laurence, S., & Margolis, E. (1997). Regress arguments against the language of thought. *Analysis*, *57*, 60–66.

[41] Laurence, S., & Margolis, E. (2001). The poverty of the stimulus argument. *British Journal for the Philosophy of Science*, *52*, 217–276. 97

[42] Leon, T., Zuccarello, P., Ayala, G., de Ves, E., & Domingo, J. (2007). Applying logistic regression to relevance feedback in image retrieval systems. *Pattern Recognition*, *40*, 2621–2632.

[43] Machery, E. (2006). Two dogmas of neo-empiricism. *Philosophy Compass*, *1/4*, 398–412. 117

[44] Margolis, E., & Laurence, S. (1999). *Concepts: Core Readings*. Cambridge: MA: MIT Press. 76, 77, 85

[45] Margolis, E., & Laurence, S. (2007). The ontology of concepts - abstract objects or mental representations? *Noûs*, *41(4)*, 561–93. 62

[46] Margolis, E., & Laurence, S. (2012). *The Oxford handbook of philosophy of cognitive science*. USA: OUP USA. 63, 64, 68

[47] Margolis, E., & Laurence, S. (2014). Concepts. *The Stanford Encyclopedia of Philosophy*. 87

[48] Medin, D., & Schaffer, M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207–238. 79

[49] Murphy, G. (2002). *The Big Book of Concepts*. Massachusetts: The MIT Press.

[50] Nosofsky, R. (1992). Exemplars, prototypes and similarity rules. *From Learning Theory to Connectionist Theory: Essays in Honor of W. K. Estes*, *1*, 149–168. 80

[51] Nosofsky, R., & Palmeri, T. (1997). An exemplar-based random walk model of speeded categorization. *Psychological Review*, *104*, 266–300. 80

[52] Ogata, K. (1987). *Sistemas Dinámicos*. México: Prentice-Hall Iberoamericana SA. 89

[53] Peacocke, C. (1992). *A Study of Concepts*. Cambridge: MA: MIT Press. 62, 64

[54] Peacocke, C. (2001). Does perception have a nonconceptual content? *The Journal of Philosophy*, *98(5)*, 239–264. 65

[55] Popper, K. (1962[1934]). *La lógica de la investigación científica*. Madrid: Tecnos.

[56] Popper, K. (1994[1963]). *Conjeturas y refutaciones*. Barcelona: Paidós Ibérica. 10

[57] Prinz, J. (2002). *Furnishing the Mind: Concepts and Their Perceptual Basis*. Cambridge: MA.: MIT Press. 96, 104, 105, 109, 111, 112, 113

[58] Prinz, J. (2005). The return of concept empiricism. In H. Cohen, & C. Leferbvre (Eds.) *Handbook of Categorization in Cognitive Science*, chap. 30, (pp. 679–694). New Jersey: Elsevier. 104

[59] Psillos, S. (2002). *Causation and Explanation*. Londres: Acumen. 10

[60] Putnam, H. (1975). The meaning of "meaning". In K. Gunderson (Ed.) *Language, mind and knowledge*, (pp. 131–192). Minneapolis: University of Minnesota Press. 82

[61] Quillian, M. R. (1968). Semantic memory. In M. Minsky (Ed.) *Semantic information processing*. Cambridge: MA.: MIT Press. 88

[62] Rolls, E., & Deco, G. (2001). *Computational neuroscience of vision*. New York: Oxford University Press. 7

[63] Roskies, A. (2008). A new argument for nonconceptual content. *Philosophy and Phenomenological Research*, *76*, 633–659. 74

[64] Rumelhart, D., McClelland, J., & the P.D.P research group (1986). *Parallel distributed processing, explorations in microstructure of cognition, 1 Foundations*. Cambridge: MIT Press. 89

[65] Russell, S., & Norvig, P. (2010). *Artificial Intelligence. A modern approach (3^a ed.)*. New Jersey: Pearson Education Inc. 12, 14, 40, 95

[66] Salmon, W., & Kitcher, P. (1989). Scientific explanation. In W. Salmon, & P. Kitcher (Eds.) *Minnesota Studies in the Philosophy of Science*, vol. 13. Minneapolis: University of Minnesota Press. 10

[67] Shepard, R. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*, 1317–1323. 80

[68] Smith, E., & Medin, D. (1981). *Categories and Concepts*. London: Harvard University Press Cambridge. 81

[69] Turing, A. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433–460.

[70] Vega de, M. (1984). *Introducción a la psicología cognitiva*. Madrid: Alianza Editorial. 79, 88

[71] Weiskopf, D. (2007). Concept empiricism and the vehicles of thought. *Journal of conciousness studies*, *14(s 9-10)*, 156–183. 117

[72] Wittgenstein, L. (1953[1958]). *Philosophical Investigations, 3rd edition*. G.E.M. Anscombe (trans.), Oxford: Blackwell. 78, 93

[73] Zalta, E. (2001). Fregean senses, modes of presentation, and concepts. *Philosophical Perspectives*, *15*, 335–359. 62

Declaration

I herewith declare that I have produced this work without the prohibited assistance of third parties and without making use of aids other than those specified; notions taken over directly or indirectly from other sources have been identified as such. This work has not previously been presented in identical or similar form to any examination board.

The dissertation work was conducted from 2008 to 2017 under the supervision of Dr. Valeriano Iranzo García at the University of Valencia.

Valencia,

This dissertation was finished writing in València on 13 de febrero de 2017

This page is intentionally left blank