Documento de trabajo

Junio 2003

Phrónimos conexionista

(Parte constructiva)

Prof. Dr. Carlos Garay

Es mi intención que exploremos los alcances, límites y consecuencias de una forma artificial, nueva, de almacenar y recuperar la información. La forma de la que estoy hablando es nueva artificialmente, pero es muy antigua biológicamente, pues se trata de la manera en la que reciben, procesan, almacenan y recuperan información los sistemas nerviosos animales. Nos ocuparemos, por supuesto, principalmente de nosotros, los seres humanos. Todos nosotros compartimos sistemas biológicos que nos han permitido incorporar, literalmente[1], una enorme cantidad de conocimientos. Y también nos han permitido generar nuevo conocimiento.

El conocimiento que acarreamos junto con nosotros lo tenemos almacenado en nuestras cabezas. A veces lo transformamos en sonidos, que llamamos palabras, para poder compartirlo con nuestros semejantes, pero esto no ocurre siempre. De hecho, no está almacenado en forma de palabras escritas ni habladas: si abrimos una cabeza nos encontraremos con carne y sangre, pero no con palabras. Nos encontramos con un conjunto de sistemas que nos permiten transformar en palabras algunas de nuestras experiencias y deseos.

Dije que compartimos sistemas biológicos, añado ahora que son muy similares, mas no idénticos. Cada uno de nosotros tenemos un rostro compuesto por las mismas partes: ojos, cejas, nariz, boca, orejas, etc., dispuestos más o menos de la misma manera. Pero resulta evidente que todos nuestros rostros son diferentes: no hay dos narices iguales, ni pares de ojos iguales. Lo mismo ocurre en nuestro interior: nuestros sistemas nerviosos están compuestos básicamente por los mismos elementos dispuestos aproximadamente de la misma manera, pero no hay dos idénticos. Y las diferencias individuales son aquí cruciales para los temas que nos ocupan. Nuestros sistemas nerviosos no sólo se diferencian entre sí por razones genéticas, como nuestras narices, sino también por razones epigenéticas, es decir, por factores que han ido produciendo esas diferencias a lo largo de nuestras vidas. Entre una persona que sabe sumar y otra que no lo sabe hay una diferencia. Pero esa diferencia no es de orden espiritual o inmaterial: es una diferencia en la composición, estructura y funcionamiento entre sus sistemas nerviosos. Esto ocurre siempre entre dos personas que tienen diferentes habilidades. De la misma manera, dos personas que tienen la misma habilidad comparten también estructuras funcionales nerviosas similares. Del hecho que ustedes hayan estudiado derecho o filosofía o física puede inferirse que sus cerebros se han ido moldeando a lo largo de los años de estudio en direcciones parecidas. Todo ese trabajo de moldeo ha sido sólo parcialmente consciente. Mucho proceso metabólico interior, absolutamente

inconsciente, fue necesario para llegar a saber lo que hoy saben: además de la asistencia a las clases y la lectura de los textos (es decir, de las experiencias específicas) fue necesario mucho alimento, sueño y entretenimiento. Hubo que darle el tiempo necesario al organismo para que asimile las experiencias específicas. La mayor parte de los procesos de aprendizaje son, pues, inconscientes y, lo que es más importante para nosotros, son de naturaleza no lingüística. Hubo una parte en la que el lenguaje jugó un papel esencial: atender las clases y leer los textos (y quizás también podamos incluir aquí las discusiones y charlas sobre esos temas). Pero una vez que el lenguaje hizo su necesaria aparición, hay que dejar el lugar correspondiente a los otros procesos sin los cuales ni ustedes ni yo sabríamos nada. Esos procesos son los que tradicionalmente se atribuyen a la inteligencia. Y esos procesos son en su gran mayoría inconscientes. Pero, además, tanto ustedes como yo no solamente aprendimos derecho o filosofía o física. También sabemos andar en bicicleta, nadar, jugar al ajedrez o tocar la guitarra. Cada una de estas habilidades también requirió un proceso interior de acomodamiento de nuestros sistemas nerviosos en los que el lenguaje desempeñó una función algo menor. Las explicaciones verbales son sin duda mucho menos importantes que la práctica en esta clase de actividades. Pero lo que más nos interesa ahora, es que además de haber aprendido derecho, filosofía y natación, hemos aprendido a comportarnos correctamente en el seno de la comunidad a la que pertenecemos. Y este es el punto al que quería llegar. Las maneras en que aprendemos los seres humanos son esencialmente las mismas: a través de ejemplos que tratamos de imitar (imitación) y a través del lenguaje que nos permite hacernos una idea (un modelo interno) de aquello que nunca hemos experimentado personalmente. El caso del comportamiento correcto no es una excepción. Hemos aprendido a comportarnos correctamente de una manera bastante parecida a la manera en que aprendimos a andar en bicicleta. Cuando estábamos en trance de aprender a andar en bicicleta y hacíamos algo mal (pedaléabamos con miedo, por ejemplo, o mirábamos nuestros pies o el manubrio) nos encontrábamos con la sanción de perder el equilibrio y caernos. Cuando nos portábamos de un modo socialmente incorrecto, también nos encontrábamos con la sanción: una reprimenda, un sopapo o un vacío social. Eso que aprendimos, andar en bicicleta o comportarnos correctamente, se encuentra almacenado en nuestros cerebros. Y si bien es cierto que, de alguna manera y a veces, podemos expresarlo en palabras, la forma en la que está almacenado es radicalmente no lingüística.

Acarreamos, pues, en nuestras cabezas, una gran cantidad de conocimientos: conocimiento de hechos de todo tipo y conocimiento de normas de todo tipo. Nuestro conocimiento del mundo, en un todo unido con las normas a que obedecen los objetos y las personas que lo pueblan, constituye para cada uno de nosotros el respaldo sobre el cual podemos elegir un curso de acción para nuestras vidas. Sabemos qué está bien hacer y qué no está bien hacer en cada caso. En la tradición aristotélica, esta capacidad para saber qué le conviene a uno en cada caso recibe el nombre de "phrónesis". La phrónesis es la virtud de juzgar lo más acertadamente posible sobre cuestiones prácticas. Y aquel que era capaz de juzgar acertadamente no sólo sobre lo que le conviene a él mismo, sino también sobre lo que le conviene a los demás, es el "phrónimos", el prudente. De alguna manera el phrónimos lleva consigo un conocimiento práctico adquirido a lo largo de la vida y como resultado de un gran número de experiencias vitales. La evolución de las civilizaciones hizo que ya no pensáramos que ese conocimiento podía ser patrimonio de personas individuales, sino más bien de organismos colegiados, concejos, parlamentos, legislaturas, etc., cuyos juicios se ejecutan por medio de otros organismos como los poderes ejecutivo y judicial, todo esto teoría de la representatividad mediante. Sin

duda, ha habido un juego histórico que va de la diversidad y complejidad creciente de las sociedades a formas cooperativas de regulación y viceversa. Sea como fuere, es imposible que una sola persona pueda concentrar en sí misma todo el conocimiento necesario para juzgar acertadamente sobre cuestiones prácticas y que este juicio sea válido para la mayoría. Asimismo, las entidades colectivas de organización, planificación y generación de políticas actúan respaldadas por el conocimiento acumulado en libros y códigos de la más diversa naturaleza.

La idea central de esta presentación sería la siguiente: si el desarrollo científico y tecnológico lo permitiera algún día, ¿podríamos volver a unificar el conocimiento práctico en un solo individuo?. Por supuesto, se trataría de un individuo artificial cuyo cerebro poseería una capacidad de procesamiento y almacenamiento de información muchísimo mayor que el de miles de personas pensando y juzgando simultáneamente.

Para que no suene tanto a ciencia ficción, propongo que examinemos un cierto cuerpo de datos y razones, que a mi juicio, hacen razonable a esta altura de la civilización comenzar a jugar con esa idea. Estos datos y razones constituirán los presupuestos sobre los que se basa la posibilidad de la idea. Muchos de ellos son datos empíricos, y por lo tanto sujetos a error y mudabilidad. Otros son especulaciones filosóficas, y por lo tanto más sujetos a error y, para colmo, sin ninguna pista que nos pueda ayudar a corregirlos. Creo que podemos otorgarle a ambos, datos y especulaciones, un provisorio voto de confianza para, de esa manera, poder formarnos una visión de conjunto.

Datos y razones de la neurobiología del comportamiento

Es bastante obvio que el principal órgano responsable del comportamiento y de la cognición es el sistema nervioso. Comprender su estructura y su funcionamiento implica, en parte, conocer cómo conocemos, es decir, cómo aprendemos, cómo almacenamos y cómo recuperamos la información necesaria para nuestra vida diaria. Nuestros sistemas nerviosos están inmersos dentro de sistemas más amplios y sólo funcionan correctamente en un entorno que les sirva de referencia, así que no será posible comprender qué es el conocimiento si nos limitamos a estudiar los sistemas nerviosos. Pero no cabe duda de que mientras no los conozcamos a fondo el conocimiento humano seguirá siendo un misterio.

Las células nerviosas se llaman neuronas. Están muy cerca unas de otras pero no llegan a tocarse. Se comunican entre sí mediante unas estructuras muy complejas denominadas "sinapsis". Cuando dos neuronas están conectadas llamamos a una "neurona presináptica" y a la otra "neurona postsináptica", unidas por el espacio sináptico o sinapsis propiamente dicha. El impulso nervioso que pasa de una neurona a otra consiste en una descarga eléctrica de la neurona presináptica originada en una diferencia de potencial eléctrico entre el exterior y el interior de la célula. Esta descarga provoca, a su vez, la descarga de la neurona postsináptica. En este proceso de pasaje de señales de una parte a otra del sistema nervioso intervienen numerosos factores. Las membranas lipídicas de las neuronas contienen "incrustaciones" de proteínas que forman canales a través de los cuales se produce el intercambio iónico. También existen en la superficie de la membrana canales receptores de sustancias (los neurotransmisores) que inhiben o facilitan ese intercambio.

Las neuronas se conectan entre sí siguiendo patrones de dos tipos: por un lado siguiendo patrones genéticos, y por otro, siguiendo patrones dependientes de la estimulación de los órganos sensoriales periféricos (ojos, oídos, piel, terminales nerviosas musculares y articulares, etc.). La manera en que se conectan las neuronas en el cerebro, la forma en la que se organizan en una arquitectura de conexiones, depende en gran parte de los patrones de estimulación recibidos por los individuos. Por ejemplo, en la corteza cerebral se pueden distinguir capas de cuerpos neuronales interconectados de una manera característica. La forma en la que se conectan esas neuronas es lo que hace funcional al sistema. Dicho de otra manera, si esas neuronas no estuvieran conectadas de la manera en la que están conectadas, el sistema no podría realizar su función. La organización final normal de un fragmento de corteza cerebral, digamos, del lóbulo occipital, permite que veamos correctamente con nuestros ojos lo que ocurre en nuestro entorno. Pero la corteza no alcanza esa organización específica si no ha sido estimulada visualmente de manera normal. Si el individuo ha sido criado en la oscuridad o si ha estado sometido a estímulos visuales anormales, la corteza no puede organizarse adecuadamente y, finalmente, no funciona: el individuo no puede ver normalmente.

Esta capacidad del sistema nervioso de modificar su estructura dependiendo de la experiencia a la que se lo somete es muy importante para nosotros. El someterse a un determinado tipo de experiencias durante la vida provoca una adaptación de las estructuras nerviosas que hace que el individuo "funcione" de cierta manera. Ser más o menos hábil para realizar una tarea de cualquier tipo (intelectual, manual, perceptual) significa haber desarrollado una arquitectura neuronal apropiada. "Aprender", en este contexto, significa haber estructurado nuestro cerebro de una manera en particular, por supuesto, en desmedro de otras posibles.

Cuando hablamos del funcionamiento de las estructuras nerviosas tenemos que destacar su aspecto dinámico. Un sistema nervioso en funcionamiento es un sistema en constante cambio. Uno de los cambios es de tipo estructural. Pero otro, no menos importante, es de tipo dinámico. Por ejemplo, mantener un número telefónico en la mente hasta el momento de discar y luego olvidarlo por completo es algo que nos ha ocurrido muchas veces. En el cerebro, mantener ese número durante cierto tiempo consiste en mantener un circuito neuronal activo con sus neuronas descargando a una determinada tasa durante ese mismo tiempo. No hay necesariamente un cambio estructural en este caso. Ni tampoco ocurre que el circuito mencionado esté completamente inactivo hasta que se lo utiliza para almacenar el número telefónico. La diferencia entre mantenerlo en mente y olvidarlo reside únicamente en la tasa de disparo de las células que intervienen en el circuito.

Así como la neurobiología nos ofrece datos acerca de la forma en la que el cerebro recibe, codifica, almacena y recupera información sobre sí mismo y sobre el entorno, la neuropsicopatología nos puede mostrar cómo ciertas alteraciones en la estructura y el funcionamiento del sistema nervioso se manifiestan en el comportamiento.

El alcoholismo crónico provoca una variada gama de alteraciones metabólicas en el organismo. Entre ellas existen algunas que causan un debilitamiento del tejido nervioso cerebral, principalmente en el lóbulo frontal, el tálamo dorsal y los cuerpos mamilares. Estos cambios en la estructura y dinámica del cerebro transforman a una persona normal en una persona con graves trastornos de memoria y de relación con

los demás. Mantener una conducta, planificarla y organizarla son tareas fuera de su alcance. Los pacientes con síndrome frontal, como los que padecen alcoholismo crónico, presentan una característica que los especialistas denominan "perseverancia". El paciente es capaz de comprender una regla como, por ejemplo, golpear las manos a cierto ritmo, pero le resulta extremadamente dificultoso cambiar de regla. En algunos casos, y dependiendo de la región del lóbulo frontal que se haya dañado, se presenta un estado de exitación en la que el sujeto toma livianamente y en broma todos los acontecimientos que le acaecen. Su lenguaje se torna grosero (algunos lo llaman eufemísticamente "coprolalia") y pierde toda noción de lo que significa comportarse de una forma socialmente aceptable. Tuve ocasión de presenciar en un hospital público el penoso espectáculo de un paciente internado que reía estrepitosamente delante de su esposa que lloraba sin consuelo. A estos pacientes les cuesta controlar sus impulsos sexuales y los originados en la ira. Pero, simultáneamente, pierden espontaneidad y no muestran interés ni por el pasado ni por el futuro. Las acciones voluntarias tienden a desaparecer: no quieren levantarse por las mañanas, no sienten la necesidad de vestirse ni de acicalarse, alimentarse ni aún la necesidad de ir al baño para satisfacer sus necesidades fisiológicas. En muchos casos muestran un comportamiento estereotipado dependiente del estímulo, por ejemplo, se ponen anteojos simplemente porque uno los puso sobre la mesa, o comen si se les presenta un planto de comida. Algunos autores contemporáneos (Spence & Frith, 1999) han interpretado que estos pacientes pierden su capacidad de elegir o iniciar una serie de actos. Pero lo más importante para nosotros es que se halla gravemente afectado su libre albedrío puesto que se vuelven esclavos de los estímulos del entorno.

Afecciones en otras áreas del cerebro provocan otras alteraciones del comportamiento pero no es adecuado exponerlas aquí debido a la extensión que llevaría. Pero me importa especialmente señalar que ya no es atendible especular acerca de la naturaleza de los actos voluntarios libres y su relación con la teoría de la culpa y del delito en términos de la grosera distinción entre locos y cuerdos o sanos y enfermos. Es indispensable el conocimiento detallado de los modos que utiliza el cerebro en la producción del pensamiento y de la acción.

Datos y razones de la neurociencia computacional

Volvamos nuestra atención ahora a los sistemas dinámicos. En primer término, nos encontraremos con un conjunto mensurable de propiedades que se va modificando con el transcurso del tiempo. Estas propiedades, en tanto mensurables, pueden representarse simbólicamente mediante cantidades por medio de conjuntos ordenados $<x_1, x_2, ..., x_n>$ en el que cada número describe el estado de la propiedad en un momento dado. A medida que el sistema evoluciona con el tiempo, las medidas van cambiando según una función, de modo que cada n-tupla corresponde a un estado del sistema en un instante dado. La secuencia de estos estados del sistema se suele llamar "espacio de estados del sistema". Así podemos decir que el sistema evoluciona dentro de su espacio de estados. (van Gelder y Port (1995) p. 7).

Según algunos autores, los sistemas dinámicos tienen la siguiente importante característica, de la que se desprenden otras dos: (1) el estado actual del sistema sólo puede determinar un único estado siguiente; (2) siendo esto así, la evolución del sistema ha de seguir alguna regla o conjunto de reglas, y (3) la sucesión de estados nunca puede bifurcarse: es decir, a partir de un cierto estado siempre se

llega al mismo tipo de estado. (van Gelder y Port, 1995, p. 6; Giunti 1995, pp. 550-551).

La manera usual de expresar las reglas de evolución del sistema incluyen el cálculo diferencial y las *difference equations*, según la descripción refleje una evolución temporal continua o discreta, respectivamente. Establecidos los parámetros del sistema es posible obtener una descripción de su desenvolvimiento, es decir, puede apreciarse cómo varían algunas de sus propiedades en función de otras.

Un subtipo de sistemas dinámicos lo constituyen las redes neurales artificiales. Una red neuronal artificial (RNA) es un conjunto de elementos interconectados. A cada unión se le asigna una cantidad, el "peso" de la unión, el cual es directamente proporcional a la facilidad con que permite el paso del impulso. Cada elemento de la red transforma los patrones de actividad que recibe de los otros elementos con los que está conectado y los transforma, mediante una función, en una única actividad de salida, la cual es transmitida a las otras unidades. El estado final de la red es una función de los impulsos que reciba, de la función de entrada-salida que se le haya asignado a las unidades y de los pesos asignados a las conexiones. Normalmente una red recibe un patrón de estímulos externos en un conjunto de unidades llamadas "unidades de entrada". Éstas transmiten el patrón de actividad recibido a las unidades de procesamiento, y allí se transforman hasta obtener un patrón de actividad de las unidades de salida. Mediante la aplicación de un tipo especial de algoritmos es posible ajustar automáticamente los pesos de las conexiones de manera que la red responda con un patrón de salida deseado a un patrón dado de entradas. Al ofrecérsele un cierto patrón de entrada, la red puede "aprender" a responder con un cierto patrón de egreso. No hay aquí símbolos ni instrucciones que formen y transformen cadenas de símbolos. Decimos que la red ya está entrenada cuando la salida se ajusta a lo que nosotros consideramos adecuado.

Una RNA es, pues, un sistema de procesamiento de la información analógico, no es digital y trabaja en paralelo. No se programa, se entrena. Consiste en varios procesadores simples, altamente interconectados, llamados nodos o también neurodos, análagos a las células nerviosas humanas, aunque tienen importantes diferencias con éstas. Los neurodos, decíamos, están conectados por un gran número de lazos de diferente peso, a través de los cuales pasa la señal. Cada neurodo recibe muchas señales de otros neurodos o del mundo externo (por ejemplo, fotones captados por una matriz de fotorreceptores, o un patrón de señales presentado a la red por el diseñador). Aunque recibe muchas señales, sólo produce una señal de output, la cual puede dividirse y alcanzar a otros neurodos. En cada una de las divisiones la señal tiene la misma intensidad. Algunos neurodos arrojan su señal de salida fuera de la red, generando de esta manera patrones de respuesta o de control.

La señal de salida de cada neurodo depende de tres factores: a) las señales de entrada, b) el peso asignado a la conexión y c) una función numérica definida para cada neurodo. Las señales de entrada se miden en magnitudes de intensidad. El peso es una cantidad que se multiplica por cada una de las señales de entrada separadamente y, sumando luego todos los productos, se obtiene la totalidad del input recibido. Por último, se procesa este número aplicándole una función que determinará, finalmente, la señal de salida del neurodo. Existen numerosas funciones que pueden cumplir este papel, pero básicamente pueden caer dentro de

tres categorías: funciones lineales, de umbral y sigmoides. En las primeras, la salida fluctúa proporcionalmente a la entrada. En las de umbral, hay dos salidas fijas: se elige una u otra según la intensidad del input total recibido supere o no cierta magnitud. Si utilizamos funciones sigmoides, la salida variará de modo no lineal según varíe la entrada.

Con sólo tres unidades de entrada que puedan tomar valores discretos de 1 a 10, tendremos una capa de entrada capaz de representar internamente 1000 estímulos diferentes.

Un cerebro humano normal contiene aproximadamente 100 mil millones (10^{11}) de neuronas. Si funcionara de manera parecida a como lo hacen las RNA, se puede ver que su capacidad de codificación sería enorme.

Uno puede entrenar una RNA para que realice determinadas tareas. Son especialmente aptas para el reconocimiento de patrones. Esto incluye: facilidad para distinguir propiedades en circunstancias inciertas, ver analogías, centrar la atención en regiones específicas del input externo, reconocer cualidades sensibles difícilmente identificables (como un aroma, o la voz de una persona). Hasta hay quienes les adjudican el poder gobernarse a sí mismas en un entorno social y moral con responsabilidad y en vistas de algún propósito.

En 1974, Paul Werbos descubrió un procedimiento matemático, conocido con el nombre de "algoritmo de retropropagación del error", que permitía ajustar automáticamente el patrón de salida de una RNA utilizando el entrenamiento mediante la exposición de la red a ejemplos. Dicho sencillamente, consiste en comparar la salida actual con el resultado deseado, y luego ir modificando sistemáticamente los pesos de las conexiones para acercarse al máximo a ese resultado. El ingeniero que diseña la red proporciona los ejemplos, y es el que determina cuándo la red ha alcanzado el resultado esperado. En este respecto, es comparable a un padre que educa a sus hijos. Él les dice qué es lo correcto tanto en el modo de pensar como en el de actuar.

Sin embargo, no es necesario que el aprendizaje de la red sea supervisado. Podemos hacer que el output deseado sea igual al input, es decir, que no sea necesario poner a disposición de la red cuál es la salida que debe tener. Obtenemos, así, un mapa del input topológicamente análogo y fiel en un sentido bastante directo.

El algoritmo de retropropagación del error es un procedimiento automático que permite calcular los pesos óptimos asignados a cada conexión utilizando como medida del error la diferencia entre el output acutal y el deseado. Necesita de, al menos, tres filas de neurodos: la primera recibe el patrón de activación proveniente del exterior de la red. La segunda "codifica en sus pesos una representación de las características presentes en los patrones de entrada", funcionando como "detectoras de características"[2]. Luego, éstas son utilizadas por los neurodos de salida para determinar el patrón correcto.

Los algoritmos utilizados en el desarrollo de RNAs son variados y responden a distintos intereses. Cuando se busca una aplicación tecnológica, como identificar huellas digitales, evaluar solicitudes de crédito o declaraciones de impuestos, los matemáticos se desentienden completamente de cuestiones biológicas y sólo hacen hincapié en la optimización operativa y en la economía del sistema. Crean así redes

sumamente eficientes pero alejadas de nuestros intereses. Es muy probable que los sistemas nerviosos funcionen de acuerdo con varias funciones de activación, aunque no sabemos cuáles. Lo importante para nosotros en este momento es que las RNAs pueden almacenar información de una manera no simbólica, que pueden establecer clasificaciones aproximadas de la información de entrada, es decir, sin utilizar necesariamente particiones estrictas sobre conjuntos de objetos, que esa información es recuperable a partir de información incompleta, y que la información se encuentra distribuida a lo largo de toda la red.

Para facilitar la exposición de estas características utilizaremos como ejemplo la RNA diseñada por Garrison Cottrell y su equipo en la Universidad de California, en San Diego[3] y que fuera entrenada para reconocer los rostros de once personas.

Está compuesta por tres capas. La hilera de entrada es una matriz de 64 X 64 unidades, cada una de las cuales admite el mismo rango continuo de niveles de brillo con topes superiores e inferiores. Su nivel de activación depende del brillo del sector de la imagen que se le presenta. La hilera intermedia (u oculta) tiene 80 neurodos, y la de salida sólo 8. De estos 8, uno se activa cuando se está en presencia de un rostro, cualquier rostro. El siguiente discrimina entre masculino o no masculino. Otro, femenino o no femenino. Los restantes 5 expresan el nombre del individuo de la imagen. Cada hilera se encuentra completamente conectada con la siguiente. El algoritmo utilizado en el entrenamiento fue el de retropropagación del error. Las imágenes utilizadas para el entrenamiento consistían en un conjunto de 64 fotografías de 11 rostros diferentes más 13 fotos de escenas que no contenían rostros. La idea era que la red fuera capaz de transformar el vector inicial de 64 X 64 elementos en uno de 80 elementos y, finalmente, en el de 8 elementos que informara correctamente si se trataba o no de una cara, el sexo y el nombre[4].

Para empezar, los pesos de las conexiones son establecidos de manera aleatoria. Al presentársele el primer input, merced a esos pesos, obtenemos una salida también caprichosa y alejada en cierto grado de la respuesta deseada. Automáticamente, la red resta el vector obtenido al vector deseado, obteniendo de esta manera la magnitud del error cometido. Luego lo eleva al cuadrado con el fin de destacar estos errores. Promediándolos, obtenemos un número que representa el tamaño del error cometido. Examinemos ahora la contribución del peso de una sola conexión a la media de error dejando fijos los demás pesos e incrementando (o decrementando) levemente el peso de esa conexión. Si el error permanece igual o aumenta, se la deja como estaba. Si, en cambio, disminuye, se fija el nuevo peso y se pasa, ordenadamente, a examinar el peso de otra conexión. Y así sucesivamente para todos los demás pesos[5]. La reiteración de este procedimiento va reduciendo gradualmente la media de error hasta un límite que depende del diseño de la red.

Este proceso puede simularse utilizando una computadora serial standard. Y existen numerosos programas que permiten estudiarlas y utilizarlas, como el *Neural Net Simmulator*, el *Rochester Connectionist Simulator*, *Stuttgart Neural Network Simulator*, etc..

Los resultados de las experiencias realizadas por Cottrell pueden resumirse así: una vez entrenada la red, se llegó a un 100% de aciertos en la identificación de los rostros que pertenecían a las fotografías utilizadas en el entrenamiento. La información que, en principio, se deseaba codificar estaba, así, correctamente almacenada en los pesos de la red. Si a estas mismas fotos se les ocultaba 1/5 del

rostro mediante una barra horizontal, se producía igualmente el reconocimiento, salvo en los casos en que la barra se ubicaba sobre la frente. En este caso, el pocentaje caía al 71%. Cuando se le presentaba otras fotografías de las personas que aparecían en el conjunto de entrenamiento pero que no se habían utilizado en el mismo, el reconocimiento se efectuó en el 98% de los casos. Y si se trataba de rostros que nunca se le habían presentado, el porcentaje de aciertos en el reconocimiento del género, bajaba al 81%.

Cada cara puede suponerse codificada por un único punto en un espacio de 80 dimensiones, correspondiendo cada una de ellas al nivel de activación de cada neurodo intermedio de la red. Para entender qué codifica cada unidad, Cottrell dedujo el patrón de activación de entrada a partir del máximo nivel de activación registrado para cada neurodo. Este patrón constituye así una suerte de fotografía generada de manera inversa, es decir, desde el estado de la red hacia lo que sería su entrada. El resultado no fue ni una parte de rostro (nariz, ojos), ni tampoco ninguno de los rostros que se le habían presentado, sino que fueron rostros "nuevos", un poco borrosos y con características mezcladas de los que habían servido de modelo. Esto puso de manifiesto que la representación de cada imagen en la red se encontraba distribuida y no localizada por partes en cada neurodo. Así puede explicarse su habilidad para el reconocimiento en condiciones de ruido, e, incluso, si se alterara o suprimiera alguno de los componentes de la red.

Si, como dijimos, a cada cara le corresponde un punto, cualquier input que lleve la actividad de la capa intermedia lo suficientemente cerca de ese punto producirá una identificación de ese input como una cara determinada. Es decir, el punto funciona como un atractor[6]. Así también, cuando se le presenta una cara que no formaba parte del conjunto de entrenamiento, el vector intermedio se parecerá lo suficiente al vector promedio de las caras originales como para identificarla como cara. Y si la cara es femenina, se parecerá suficientemente al promedio de las caras femeninas como para identificarla como una cara femenina, y así sucesivamente. El espacio de representación se encuentra, pues, dividido en sectores, cada uno de los cuales representa una categoría.

La formación de estas categorías depende de las entradas que hayamos ofrecido a la red. En el caso del ejemplo han sido fotografías de rostros, pero, al menos en principio, podría tratarse de cualquier clase de input físico del que puedan obtenerse los parámetros relevantes para codificar la entrada.

Paul Churchland tiene la osadía de llamar "conceptos" a estas particiones. Y llega a describir la capacidad de las RNAs de recuperar información a partir de una entrada incompleta como una inducción, es decir, como una tendencia de la red a saltar a una conclusión contando sólo con una evidencia parcial.

En los casos de mayor éxito, en el ejemplo, reconocimiento de los rostros particulares del conjunto de entrenamiento, la correspondencia es total. La red ha dispuesto de los recursos suficientes para ejecutar completamente la tarea de reconocimiento de esas fotografías. Esos rostros están completamente representados en la red y, por lo tanto, la representación es verdadera. Dado el mecanismo codificador, el input ha causado esta representación, teniendo de esta manera una concepción clara de cómo se ha producido y en qué consiste la correspondencia entre el input y el output.

La activación de una neurona biológica individual depende, en primera instancia, de la diferencia de potencial eléctrico entre el interior y el exterior de la célula. La diferencia aumenta o disminuye de acuerdo con la actividad electroquímica de las neuronas adyacentes que estén en contacto sináptico con ella. Parece, pues, bastante natural describir su grado de activación como una función del grado de activación de las células adyacentes. La evolución de la respuesta neuronal puede describirse utilizando los recursos matemáticos de la dinámica de sistemas.

Resumen

He querido mostrar, mediante breves indicaciones y ejemplos, que lo que sabemos acerca del comportamiento humano y sus bases físicas sumado al desarrollo de nuevas herramientas matemáticas y sus correlatos tecnológicos hacen, al menos, plausible la idea de crear una conciencia práctica, es decir, una inteligencia artificial que no solucione solamente problemas de cálculo, sino también asuntos en los que intervienen valoraciones y ponderaciones no accesibles a la inteligencia artificial clásica, reducida como está ahora al desarrollo de algoritmos para la manipulación de símbolos.

He omitido muchos puntos de interés con el fin de expresar la idea con la mayor claridad posible. Por ello, quisiera remarcar lo siguiente:

a) Tenemos la capacidad de juzgar rectamente sobre múltiples cuestiones.

b) En tanto individuos, somos muy falibles. Históricamente hemos llegado a tomar decisiones basadas en el voto popular y en las deliberaciones de cuerpos colegiados.

c) Cada sistema nervioso puede entenderse como una unidad de procesamiento parcialmente autónoma con un sistema propio de almacenamiento analógico de la información.

d) El phrónimos conexionista sería una unidad de procesamiento artificial parcialmente autónoma cuyo sistema de almacenamiento de la información sea el resultado de la integración del contenido de los sistemas individuales (probablemente sólo de algunos sistemas individuales) mencionados en el punto anterior. Sería un individuo artificial que está en condiciones representar a muchos individuos naturales.

La Plata, junio de 2003

ANEXO

Wolfgang Maas[7] nos ofrece una concisa definición de "arquitectura de red neural" en términos matemáticos. Según Maas, una red neural N es un tipo especial

de grafo. Sus nodos de entrada y de salida se encuentran rotulados (*labeled*) por números naturales. Un nodo ***g*** en *N* con una entrada ***r*** > 0 se llama *nodo de computación* que se identifica mediante alguna función de activación g_g: R ® R, algún polinomio Q_g (y_1, ..., y_r) y un subconjunto P_g compuesto por coeficientes de ese polinomio. Para el caso, P_g podría estar compuesto por todos los coeficientes. Cada elemento de P_g, recibe el nombre de *parámetro programable* de *N* . Los coeficientes que no pertenecieran a P_g serán los *parámetros no programables*. Ahora bien, si *N* tiene ***w*** parámetros programables, y se dan las siguientes condiciones: se asignaron valores a todos los no programables, se fijaron algunos programables, y *N* tiene *d* nodos de entrada y *l* nodos de salida, entonces, cada asignación a Î R^w a los parámetros programables de *N* define un circuito analógico $N^{\underline{a}}$, el cual computa una función $\underline{x}$ $N^{\underline{a}}(x)$ de R^d en R^l de la manera siguiente: supongamos que alguna entrada $\underline{x}$ Î R^d ha sido asignada a los nodos de entrada. Si un nodo ***g*** en *N* tiene ***r*** predecesores inmediatos cuya salida es y_1, ..., y_r Î R, entonces, la salida de ***g*** será g_g (Q_g (y_1, ..., y_r)).

---

[1] La palabra "incorporar" no es casual ni gratuita en este caso, sino tomada en su sentido etimológico de "meter dentro del cuerpo".

[2]Caudill, M. y Butler, Ch., *Understanding Neural Networks*, MIT Press, 1994, p. 173.

[3]Cottrell, G., "Extractig features from faces using compression networks: Face, identity, emotions and gender recognition using holons", en Touretzky, D. et al. (eds.) *Connectionist Models: Proceedings of the 1990 Summer School*, Morgan Kaufmann, San Mateo, California, 1991. También encontramos una buena descripción de su trabajo en Paul Churchland, *The engine of reason, the seat of soul*, MIT Press, 1996, cap. 3.

[4]El nombre es acá solamente un código arbitrario de 5 dígitos.

[5]He tomado esta descripción simple del procedimiento de Paul Churchland (1996), *op. cit.*, pp. 43-44.

[6]Técnicamente hablando, *A* es un atractor si es un subconjunto del espacio de estados tal que 1) la trayectoria de cualquier punto en *A* permanece en *A*, es decir, *A* es invariante; 2) los puntos que se acercan lo suficiente a *A*, tienden a acercarse cada vez más a *A* a medida que pasa el tiempo, y 3) *A* no contiene subconjuntos cerrados más pequeños con las propiedades 1) y 2). Si *d(x,A)* denota la distancia entre *x* y el

conjunto *A*, entonces, la condición 2) puede expresarse así: existe un número *d > 0* tal que si *d(x,A) < d*, entonces, en el caso de un flujo j, *lim d(j(t,x),A) = 0*, para *t* tendiendo a infinito.

[7]Maas, Wolfgang "Perspectives of Current Research about the Complexity of Learning on Neural Nets", en V. P. Roychowdhury et al. (eds.), *Theoretical Advances in Neural Computation and Learning*, Kluwer Academic Publishers, en prensa.