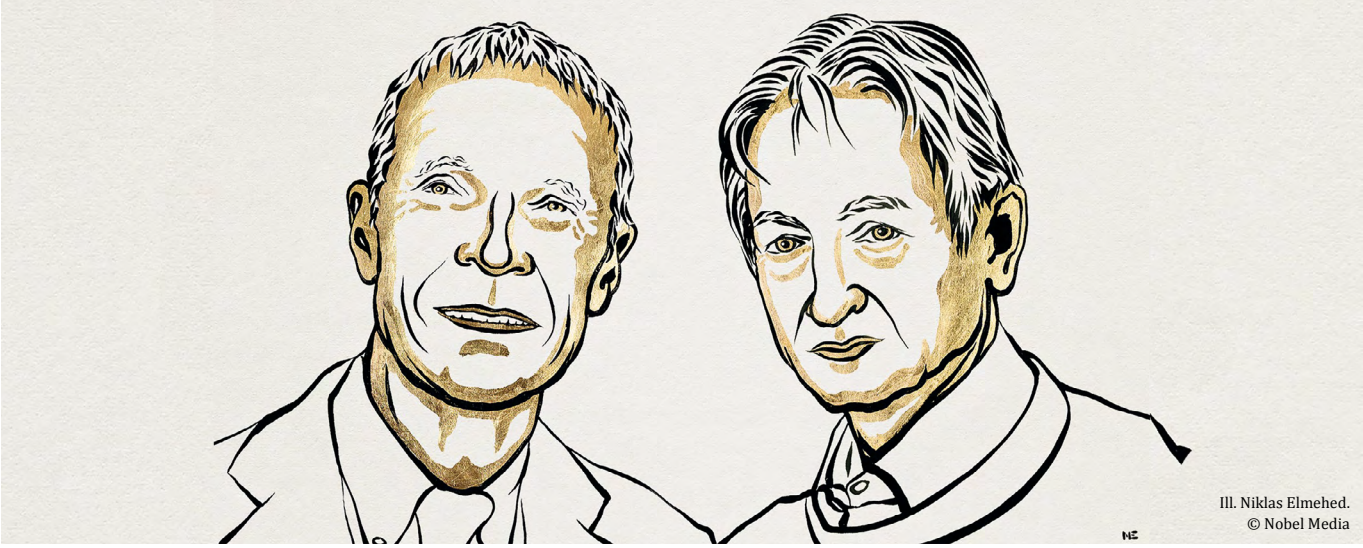
COMENTARIO INVITADO

Premio Nobel de Física 2024

Descubrimientos e inventos fundamentales que impulsan el aprendizaje automático con redes neuronales artificiales

Néstor Parga



Ill. Niklas Elmehed.
© Nobel Media

El Premio Nobel de Física del año 2024 marca un hito fundamental en la historia de la ciencia, ya que, por primera vez, se otorgó a avances en redes neuronales, disciplina que impulsa el aprendizaje automático y la inteligencia artificial en la actualidad. Este reconocimiento pone de manifiesto el profundo impacto de este campo, destacando su papel esencial en la transformación tanto de los marcos teóricos como de las aplicaciones prácticas en la ciencia contemporánea.

Reconocimiento a las redes neuronales

John J. Hopfield y Geoffrey Hinton recibieron el premio Nobel de Física por sus contribuciones al desarrollo de las redes neuronales artificiales. El modelo de memoria asociativa propuesto por Hopfield [1] reveló que el almacenamiento y la recuperación de información en redes neuronales pueden surgir como propiedades emergentes de su dinámica. El modelo se inspiró en conceptos de sistemas magnéticos, estableciendo una analogía entre el aprendizaje en redes neuronales y la dinámica colectiva de espines interactuantes.

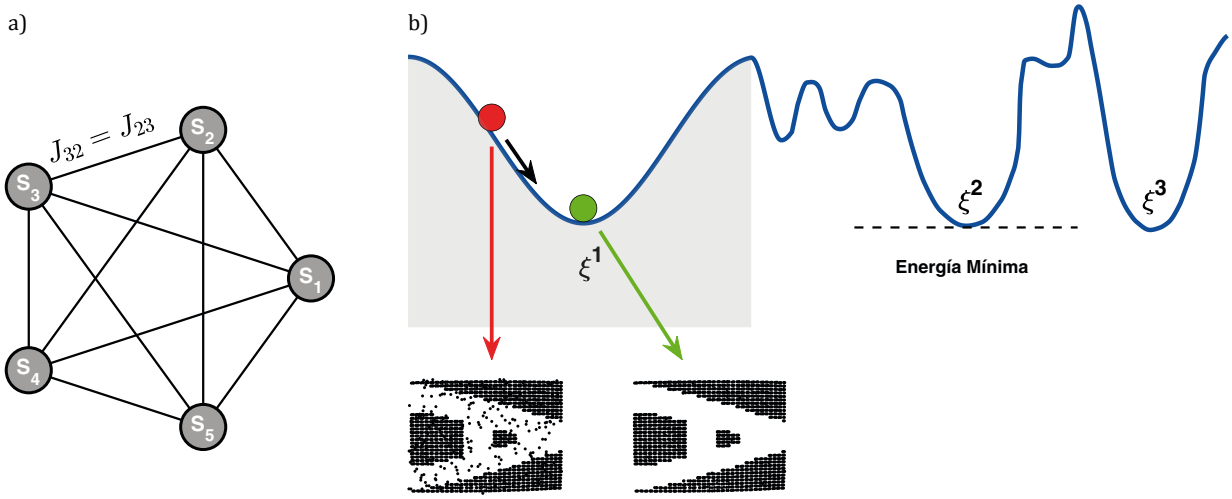


Fig. 1. (a) Red de Hopfield. S_i indica el estado de la neurona i-ésima a un tiempo fijo de la dinámica de la red. J_{ij} es el acoplamiento simétrico entre las unidades i y j, que se construye en términos de las P memorias almacenadas en la red. **(b)** Memoria asociativa en una red de Hopfield. En la parte superior, el paisaje de energía, donde se almacenan tres patrones, por ejemplo, representaciones de la letra A y otras dos letras, utilizando círculos y espacios en blanco. En la parte inferior: a la derecha, la representación del patrón A; a la izquierda, una versión ruidosa del mismo con un 10 % de ruido. La red, al ser inicializada con una versión alterada del patrón, evoluciona hasta recuperar completamente el patrón memorizado A, corrigiendo así los errores. La red posee estados de mayor energía. La zona en gris indica la cuenca de atracción del patrón A.

Modelo de Hopfield

El modelo de Hopfield [1] es una red neuronal recurrente que almacena P patrones de actividad ξ_i^μ, con $i = 1, \ldots, N$ representando las N neuronas de la red, y $\mu = 1, \ldots, P$ indicando los distintos patrones. Cada componente espacial ξ_i^μ es una variable binaria que toma los valores $\xi_i^\mu \in \{-1,1\}$. El estado de actividad de la neurona i en el tiempo t se denota como $S_i(t)$, con $S_i(t) \in \{-1,1\}$. La actualización del estado de cada neurona sigue la siguiente regla

$$S_i(t+1) = \mathrm{sgn}\left(\sum_{j=1}^{N} J_{ij} S_j(t)\right),$$

donde J_{ij} son las conexiones sinápticas entre las neuronas i y j, y $\mathrm{sgn}(x)$ toma los valores 1 si $x > 0$ y –1 si $x < 0$. Los patrones de actividad ξ_i^μ se almacenan en la red mediante una regla hebbiana para las sinapsis (se asume $J_{ij} = 0$ para evitar autoconexiones)

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu.$$

Para demostrar que en el límite $N \to \infty$ los patrones de actividad ξ_i^μ son atractores de la dinámica, supongamos que el estado inicial $S_i(0)$ está cerca de uno de los patrones, ξ_i^ν. Entonces $S_i(t+1)$ es

$$S_i(t+1) = \mathrm{sgn}\left(\sum_{j=1}^{N} \sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu S_j(t)\right).$$

En el tiempo, t $S_i(t) \approx \xi_i^\nu$. La expresión entre los paréntesis se puede descomponer

$$\frac{\xi_i^\nu}{N} \sum_{j=1}^{N} \xi_i^\nu S_j(t) + \sum_{\mu \neq \nu} \frac{\xi_i^\mu}{N} \sum_{j=1}^{N} \xi_j^\mu S_j(t)\,.$$

El primer término corresponde al patrón correcto $\mu = \nu$, mientras que el segundo representa interferencia de los otros patrones. Consideremos el límite de $N \to \infty$. En este límite, la interferencia de los patrones $\mu \neq \nu$ tiende a cero, ya que los patrones ξ_i^μ son independientes y no correlacionados entre sí. Por lo tanto, en el límite de $N \to \infty$ la dinámica del modelo de Hopfield hace que el estado de la red converja al patrón ξ_i^ν, lo que demuestra que los patrones almacenados son atractores de la dinámica.

La dinámica del modelo de Hopfield puede interpretarse como la evolución a temperatura cero de un sistema de física estadística. En este contexto, el sistema minimiza una función de energía (o hamiltoniano) definida como:

$$E = -\frac{1}{2} \sum_{i,j} J_{ij} S_i S_j\,.$$

La actualización de cada neurona corresponde a un proceso de descenso de energía, ya que la regla de actualización $S_i(t+1) = \mathrm{sgn}\left(\sum_j J_{ij} S_j(t)\right)$ reduce el valor de E en cada paso de la dinámica. Los atractores corresponden a los mínimos locales de E, que son precisamente los patrones almacenados ξ_i^μ (figura 1b).

Esta relación resalta el vínculo profundo entre la física estadística y la neurociencia computacional. Por otro lado, la máquina de Boltzmann, desarrollada por Hinton y colaboradores [2], aplica principios probabilísticos de la física estadística para habilitar el aprendizaje en redes neuronales. Este enfoque supuso un avance crucial para el aprendizaje automático, facilitando la detección de patrones en datos mediante aprendizaje no supervisado. Los trabajos de Hopfield y Hinton unieron conceptos de la mecánica estadística con la neurociencia computacional y el aprendizaje automático, proporcionando un marco teórico que influiría profundamente en estas disciplinas y en muchas otras áreas de la ciencia y de la tecnología.

Hopfield y la memoria asociativa

En su trabajo de 1982, Hopfield propuso un modelo de red neuronal recurrente compuesto por simples neuronas binarias [1], con una dinámica gobernada por un principio de minimización de energía (figura 1a; véase el recuadro Modelo de Hopfield). Su gran contribución radicó en mostrar que estas redes podían almacenar múltiples patrones de memoria.

Bajo determinadas condiciones, los patrones almacenados actúan como atractores de la dinámica del modelo, es decir, como puntos fijos estables. El estado inicial de la red determina hacia qué atractor converge su evolución. Cada atractor posee una cuenca de atracción, definida como el conjunto de estados iniciales que conducen a dicho patrón (figura 1b). Debido a que un patrón almacenado puede ser recuperado a partir de una versión incompleta o ruidosa del mismo, se dice que la memoria es de tipo autoasociativa.

La solución de la termodinámica del modelo de Hopfield [3] permitió establecer las condiciones bajo las cuales los patrones almacenados son atractores estables de la dinámica del sistema, en función de la temperatura $T = 1/\beta$ y del parámetro de carga $\alpha = P/N$, donde N es el número de neuronas y P es el número de patrones almacenados. Para valores bajos de la temperatura T y una carga α pequeña, los patrones almacenados $\{\xi_i^\mu\}$ ($\mu = 1, \ldots, P$) son atractores estables (véase el recuadro Modelo de Hopfield). En esta fase, la actividad de las neuronas se alinea con un patrón μ. Su valor medio $\langle S_i \rangle$ cumple

$$\frac{1}{N} \sum_i \langle S_i \rangle \xi_i^\mu = M, \qquad \frac{1}{N} \sum_i \langle S_i \rangle \xi_i^\nu = 0 \quad (\nu \neq \mu),$$

donde M depende de α y β. Existen dos estados de recuperación, uno con $M > 0$ y otro con $M < 0$. En esta situación, la red neuronal puede recuperar de manera confiable un patrón almacenado incluso a partir de una versión parcial o ruidosa, gracias a la presencia de una cuenca de atracción suficientemente amplia alrededor de cada patrón.

Esta fase es conocida como la fase de recuperación (figura 2).

Los estados de recuperación emergen de forma metastable por debajo de $T_M(\alpha)$ y se convierten en mínimos globales estables por debajo de $T_c(\alpha)$, lo que marca una transición de primer orden. En esta región intermedia entre $T_c(\alpha)$ y $T_M(\alpha)$ aparecen estados que son configuraciones de la red que no corresponden exactamente a patrones almacenados, pero que tienen una energía relativamente baja y pueden atrapar transitoriamente la dinámica de la red. Estos estados están relacionados con configuraciones cercanas a los patrones almacenados, pero no son globalmente estables. A valores altos de α o T, entre $T_M(\alpha)$ y $T_g(\alpha)$, el sistema entra en la fase de vidrio de espín (figura 2). En esta fase, los estados de la red no están correlacionados con los patrones almacenados. Los estados de vidrio de espín son configuraciones de energía baja pero desordenadas, donde las neuronas presentan una polarización espontánea que no refleja ninguna estructura de los patrones almacenados. Este comportamiento es caracterizado por un parámetro de orden α, que mide la magnitud del desorden interno en la red, y no hay alineamiento con ningún patrón [4]

$$q = \frac{1}{N}\sum_i \langle S_i\rangle^2 > 0, \quad \frac{1}{N}\sum_i \langle S_i\rangle\, \xi_i^\mu = 0 \quad \forall \mu\,.$$

Finalmente, para temperaturas altas, $T > T_g(\alpha)$, el sistema se encuentra en una fase paramagnética, donde las neuronas no presentan polarización espontánea, es decir, $\langle S_i\rangle = 0$ (figura 2). El diagrama de fases refleja la naturaleza rica y compleja del modelo de Hopfield, que combina propiedades emergentes propias de sistemas físicos con capacidades computacionales relacionadas con el almacenamiento y recuperación de información [5].

Tras la aparición del artículo de Hopfield, publicado en la sección de biofísica de la revista *PNAS* [1], las reacciones se extendieron en varias áreas de la ciencia. En neurociencia, el trabajo de Hopfield surgió en un momento de cambio de paradigma. La tradicional {doctrina de una neurona, que había dominado desde los estudios de Ramón y Cajal [6], se basaba en la idea de que las capacidades cognitivas del cerebro eran atribuibles a neuronas individuales. Sin embargo, una nueva perspectiva comenzaba a extenderse: la doctrina de los circuitos cerebrales, según la cual las propiedades cognitivas del cerebro no dependen de neuronas aisladas, sino de las interacciones dentro de poblaciones neuronales [7]. El modelo de Hopfield representó un avance crucial hacia esta visión, al demostrar que una red neuronal podía recuperar información almacenada en las conexiones sinápticas entre neuronas, proporcionando un modelo matemático robusto de la memoria asociativa. Por otra parte, la presencia de atractores se utilizó para proponer modelos en los que la información podría mantenerse en memoria en la forma de actividad neuronal persistente durante tiempos del orden de segundos. Aunque este no es el único mecanismo cerebral que podría dar lugar a la memoria de corto plazo, esta propuesta ha recibido una enorme atención tanto conceptual como experimentalmente [8].

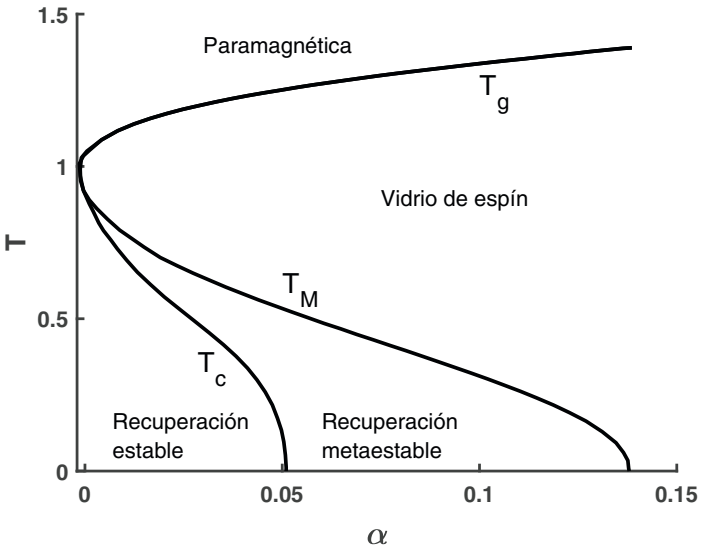


Fig. 2. Diagrama de fases de la red de Hopfield. La línea T_g marca la frontera entre la fase paramagnética y la fase de vidrio de espín. A la temperatura T_M emerge la fase de recuperación de memorias, la cual, para $T < T_c$, se convierte en el mínimo global del sistema.

En inteligencia artificial contribuyó a focalizar la atención en redes neuronales artificiales con arquitectura recurrente, en la que todas las unidades que componen la red pueden conectarse entre sí. También motivó un enfoque de la inteligencia artificial basado en la existencia de un hamiltoniano. Un pronto ejemplo de eso fue la máquina de Boltzmann, propuesta por Hinton y colaboradores [2] y, más recientemente, el modelo de memoria asociativa densa, que define redes con enorme capacidad de almacenamiento [9,10].

En el ámbito de la mecánica estadística, donde la emergencia de propiedades macroscópicas en sistemas de partículas interactuantes era un concepto bien establecido, la semejanza entre el modelo de Hopfield y sistemas magnéticos desordenados, particularmente con los vidrios de espín [4], motivó, especialmente desde mediados de la década de 1980, a numerosos físicos a investigar las propiedades de redes neuronales, lo que dio impulso a la neurociencia computacional [5,11,12]. Desde esta disciplina se propusieron soluciones para resolver algunas limitaciones del modelo original.

En el modelo de Hopfield, aprender nuevos patrones más allá de la capacidad máxima destruye la estabilidad de las memorias previamente almacenadas, impidiendo su recuperación. Para resolver este problema, Giorgio Parisi propuso limitar el rango de las sinapsis [13]; la red resultante puede operar en un régimen en el que las memorias antiguas decaen gradualmente al incorporar nuevas. Esto evita la eliminación catastrófica por interferencia, aunque la capacidad del modelo se reduce.

El modelo de Hopfield emplea una regla de aprendizaje efectiva para almacenar patrones ortogonales. Sin embargo, en el mundo real, el conocimiento se organiza de manera jerárquica, en lugar de conformar conjuntos de patrones aislados e independientes. Para abordar esta limitación, Parga y Virasoro ampliaron el modelo de Hopfield, permitiéndole almacenar un árbol jerárquico completo de categorías junto con sus subcategorías y relaciones. A través de herramientas de la mecánica estadística, demostraron que la categorización surge de manera natural cuando la memoria asociativa incorpora un proceso de codificación en capas y una regla de aprendizaje adaptada a esta estructura [14].

La estrecha relación entre el modelo de Hopfield y los modelos de vidrios magnéticos es evidente en el libro *Spin Glass and Beyond*, de Giorgio Parisi y colaboradores [4], en el cual aparece reimpreso el artículo de Hopfield de 1982, así como también algunos de los artículos mencionados [3,14].

La consideración de redes redes recurrentes en neurociencia fue mucho más allá de proporcionar modelos de memoria asociativa. Redes con acoplamientos aleatorios condujeron a la explicación de fenómenos de la actividad cortical tales como la irregularidad de los disparos de las neuronas en los circuitos corticales [15] y su asincronicidad [16]. En la última década, la posibilidad de entrenar redes neuronales en tareas cognitivas, similares a las utilizadas en laboratorios de electrofisiología, proporcionaron modelos capaces de generar hipótesis sobre cómo el cerebro resuelve esas tareas [17,18,19].

Hopfield y la memoria asociativa densa

En el modelo de Hopfield, cuando el número de memorias almacenadas es considerablemente mayor que el número de neuronas, la red neuronal entra en una fase de vidrio de espín caracterizada por la presencia de mínimos locales que no guardan correlación con los vectores de memoria (figura 2). Esto implica una baja capacidad del modelo, que solo crece linealmente con el número de neuronas. La cuestión de cómo conseguir un número de estados estables que aumente con N de manera supralineal y que al mismo tiempo tengan cuencas de atracción grandes no se resolvió hasta muy recientemente, con el modelo moderno de Hopfield [9,10]. Como en estos nuevos modelos los patrones almacenados aparecen más densamente que en la red tradicional de Hopfield también se los denomina modelos de memoria asociativa densa (MAD). Se trata de una familia de modelos que generalizan a la red de Hopfield reemplazando la interacción cuadrática en la función energía por una interacción de mayor orden o exponencial. Más específicamente,

$$E = -\sum_{\mu=1}^{P} F\left(\sum_{i=1}^{N} \xi_i S_i\right),$$

donde $F(x)$ es la función de interacción. Krotov y Hopfield propusieron $F(x) = x_n$ [9], obteniendo una capacidad $P_{\max} \sim N^{n-1}$ ($n > 2$). Cabe notar que la red de Hopfield tradicional corresponde a $n = 2$. Se puede obtener una capacidad aun mayor con una interacción exponencial, $F(x) = \exp(x)$ [10], en cuyo caso $P_{\max} \sim \exp(aN)$, con $a < \ln 2/2$. El resultado es sorprendente porque intuitivamente una capacidad grande se asocia con cuencas de atracción pequeñas. Sin embargo, en el caso exponencial, el radio de la esfera que contiene estados que siguiendo la dinámica del modelo convergen a la memoria correcta es similar al de la red de Hopfield [10].

De manera inesperada, el modelo moderno de Hopfield está estrechamente relacionado con un avance fundamental en inteligencia artificial. Recientemente se introdujo en el aprendizaje automático una nueva arquitectura de red neuronal que incorpora un mecanismo de atención y que ha tenido un gran impacto en el procesamiento del lenguaje: el transformador (*transformer*) [20]. Poco después de esta innovación, se propuso una variante del modelo moderno de Hopfield, que utiliza variables continuas y cuya regla de actualización corresponde al mecanismo de atención del transformador, lo que ha impulsado el desarrollo de nuevas arquitecturas profundas [21]. El hecho de que el MAD esté basado en un hamiltoniano es de gran relevancia, ya que introduce la

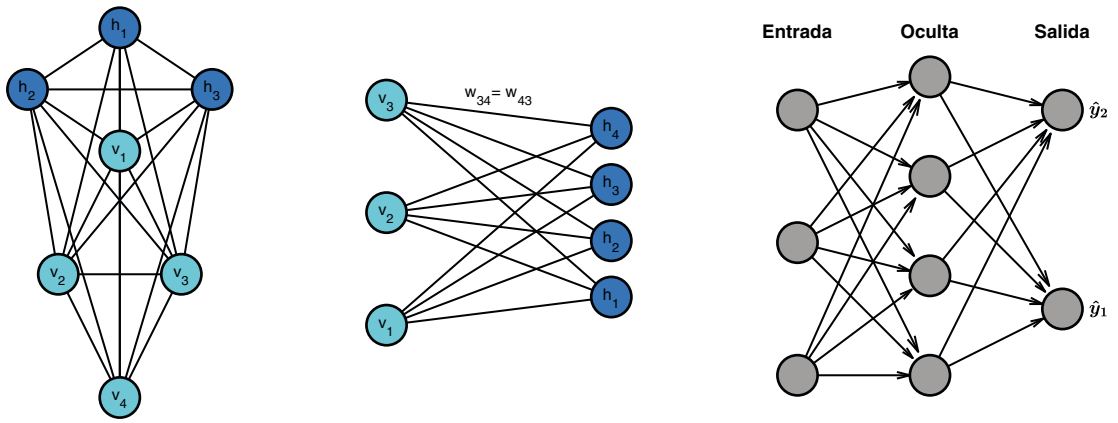


Fig. 3. (a) Máquina de Boltzmann con unidades visibles (azul claro) y ocultas (azul oscuro). Las líneas indican las conexiones simétricas entre las unidades. **(b)** Máquina de Boltzmann Restringida sin sesgos. Se emplean las mismas convenciones que en el panel a. **(c)** Arquitectura de red de avance directo (red *feedforward*) con una única capa de neuronas ocultas. Las dos salidas de esta red se indican con $\hat{y}_1$ e $\hat{y}_2$.

posibilidad de interpretar las reglas que definen los transformadores en términos de principios físicos establecidos, como la optimización de una función de energía, lo que podría facilitar el diseño de nuevas arquitecturas más eficientes y teóricamente fundamentadas.

Hinton y la máquina de Boltzmann

El modelo de Hopfield tiene varias limitaciones, incluyendo su carácter determinista y que una vez que ha sido preparado para almacenar un conjunto de patrones, los pesos se mantienen constantes. Estas características restringen su aplicabilidad a problemas que requieren manejar incertidumbre, aprender representaciones más complejas, o ajustarse gradualmente a nuevos datos. Para superar estas limitaciones, Hinton y Sejnowski introdujeron la máquina de Boltzmann (MB) como una extensión del modelo de Hopfield, incorporando estocasticidad y actualización gradual de pesos [2].

La MB es una red neuronal recurrente diseñada para modelar distribuciones de probabilidad. Es un modelo estocástico que consta de dos tipos de unidades: visibles y ocultas. Las unidades visibles representan las variables observables del sistema (datos), mientras que las ocultas modelan características latentes no directamente accesibles. Estas últimas permiten que la MB aprenda representaciones internas más ricas y capte dependencias complejas en los datos. Las unidades visibles se denotan como v_i y las ocultas como h_j ($i = 1, \ldots, N_v$ y $j = 1, \ldots, N_h$, respectivamente) y el estado de la red se indica por $(\mathbf{v},\mathbf{h})$. Todas las unidades tienen estados binarios v_i, $h_j \in \{0, 1\}$, y están conectadas mediante pesos simétricos (figura 3a). En el recuadro Máquina de Boltzmann se describe una arquitectura simplificada de MB (figura 3b). Cada unidad tiene una probabilidad de activarse determinada por una función de energía $E(\mathbf{v},\mathbf{h})$, donde la estocasticidad del modelo facilita la exploración de configuraciones múltiples y captura la incertidumbre en los patrones. A diferencia del modelo de Hopfield, la MB no solo sirve para la recuperación de memoria, sino también para el aprendizaje generativo y la clasificación. En tareas generativas, la MB aprende una distribución $P(\mathbf{v})$, permitiendo generar ejemplos nuevos mediante muestreo. En clasificación, parte de las unidades visibles se condicionan a las etiquetas de clase, de modo que el modelo aprende la probabilidad condicional $P(\text{clase}|\text{datos})$. Gracias a su estructura estocástica y la inclusión de unidades ocultas, la MB ofrece mayor flexibilidad y poder expresivo, aplicándose en un rango más amplio de tareas en comparación con el modelo de Hopfield.

En una MB, los pesos w_{ij} se actualizan gradualmente siguiendo una regla de aprendizaje basada en gradientes, con el objetivo de minimizar la energía promedio del sistema y ajustar la distribución modelada $P(\mathbf{v},\mathbf{h})$ a los datos observados. El

Máquina de Boltzmann

La máquina de Boltzmann restringida (MBR) es una red de dos capas en la que solo hay conexiones entre unidades visibles y ocultas (figura 3b).

Su función de energía $E(\mathbf{v},\mathbf{h})$ para una configuración específica $(\mathbf{v},\mathbf{h})$ es

$$E(\mathbf{v},\mathbf{h}) = -\sum_{i,j} w_{ij} v_i h_j - \sum_i b_i v_i - \sum_j c_j h_j ,$$

donde w_{ij} es el peso entre la unidad visible i y la oculta j; b_i y c_j son sesgos asociados a esas unidades, respectivamente.

La probabilidad conjunta de $(\mathbf{v},\mathbf{h})$ está dada por la distribución de Boltzmann

$$P(\mathbf{v},\mathbf{h}) = \frac{e^{-E(\mathbf{v},\mathbf{h})}}{Z} , \qquad Z = \sum_{\mathbf{v},\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})},$$

donde Z es la *función de partición*.

La probabilidad marginal de los estados visibles, que es relevante para los datos observables (la *imagen binaria* $\mathbf{v}$), es

$$P(\mathbf{v}) = \sum_{\mathbf{h}} P(\mathbf{v},\mathbf{h}) .$$

La probabilidad de que la red asigna a una imagen de entrenamiento puede aumentarse ajustando los pesos y sesgos para reducir su energía y aumentar la de otras, especialmente aquellas con energías bajas, ya que contribuyen significativamente a Z. El aprendizaje ajusta los pesos w_{ij} para maximizar la probabilidad de los datos con el método de gradiente. La derivada de $P(\mathbf{v})$ con respecto al peso w_{ij} es

$$\frac{\partial \log P(\mathbf{v})}{\partial w_{ij}} = \langle v_i h_j \rangle_{\text{data}} - \langle v_i h_j \rangle_{\text{model}}$$

que nos da la regla de aprendizaje

$$\Delta w_{ij} = \eta(\langle v_i h_j \rangle_{\text{data}} - \langle v_i h_j \rangle_{\text{model}}),$$

siendo $\langle . \rangle_{\text{data}}$ y $\langle . \rangle_{\text{model}}$ valores medios tomados con los datos o con el modelo; η es la tasa de aprendizaje. Como en una MBR no existen conexiones directas ni entre las unidades ocultas ni entre las unidades visibles, es muy sencillo obtener una muestra de $\langle v_i h_j \rangle_{\text{data}}$.

Dada una imagen de entrenamiento seleccionada aleatoriamente, $\mathbf{v}$, el estado binario h_j de cada unidad oculta j se establece en 1 con probabilidad

$$P(h_j = 1|\mathbf{v}) = \sigma\left(\sum_i w_{ij} v_i + c_j\right) ,$$

donde $\sigma(x) = \frac{1}{1+e^{-x}}$ es la función sigmoide.

También es muy fácil obtener una muestra del estado de una unidad visible, dado un vector oculto $\mathbf{h}$

$$P(v_i = 1|\mathbf{h}) = \sigma\left(\sum_j w_{ij} h_j + b_i\right).$$

Obtener una muestra de $\langle v_i h_j \rangle_{\text{model}}$ es mucho más complicado. Para lograrlo, se puede iniciar desde un estado aleatorio de las unidades visibles y ejecutar un muestreo de Gibbs alternante durante un tiempo prolongado. No obstante, Hinton introdujo un procedimiento mucho más rápido [21]. La optimización de los sesgos $\mathbf{b}$ y $\mathbf{c}$ es similar.

entrenamiento de la MB, tal como fue concebida inicialmente por Hinton y Sejnowski [2], enfrenta serias dificultades computacionales debido a la necesidad de realizar muestreo para implementar la regla de aprendizaje basada en gradientes. Este proceso requiere aproximar la distribución de probabilidad conjunta $P(\mathbf{v},\mathbf{h})$ mediante métodos como el *simulated annealing* (recocido simulado), un algoritmo estocástico que explora el espacio de configuraciones reduciendo gradualmente la temperatura del sistema. Sin embargo, este método es extremadamente lento, ya que puede requerir un gran número de iteraciones para converger a una solución cercana al equilibrio. Este coste computacional limita significativamente la escalabilidad del modelo y su aplicación práctica en problemas grandes.

Para superar estas limitaciones, se introdujo la máquina de Boltzmann restringida (MBR) [21]. La MBR simplifica la estructura de la red original eliminando las conexiones entre las unidades de un mismo tipo (figura 3b), lo que facilita su entrenamiento (véase el recuadro Máquina de Boltzmann). Esta restricción reduce drásticamente la complejidad computacional del muestreo, ya que las unidades visibles y ocultas son condicionalmente independientes entre sí dado el estado de las otras. Las MBR han demostrado ser especialmente útiles en el preentrenamiento de redes neuronales profundas [22], facilitando el aprendizaje.

Adicionalmente, Hinton introdujo la regla de aprendizaje llamada *divergencia contrastiva* [21], que proporciona una estimación más eficiente al gradiente de aprendizaje. En lugar de esperar a que el muestreo alcance el equilibrio, la divergencia contrastiva utiliza una cantidad fija y pequeña de pasos de muestreo de Gibbs para estimar la distribución posterior, reduciendo drásticamente el tiempo de entrenamiento sin sacrificar demasiado la calidad del aprendizaje. Aunque esta técnica no garantiza una estimación precisa del gradiente, en la práctica se ha mostrado efectiva para entrenar la MBR y aprender representaciones útiles.

En resumen, la MBR resuelve el problema de la ineficiencia en el muestreo al simplificar la estructura del modelo, mientras que la divergencia contrastiva acelera el aprendizaje aproximando el gradiente de manera computacionalmente eficiente. Estas innovaciones han permitido la amplia aplicación de MBR en diversos problemas de aprendizaje automático.

Hinton y la retropropagación de errores

En la contribución inicial de Hinton y colaboradores, se estableció formalmente el marco de retropropagación (*backpropagation*) [23]. El objetivo del algoritmo es minimizar por el método de gradiente una función de coste $C(\hat{\mathbf{y}}, \mathbf{y})$, que mide la diferencia entre las salidas predichas $\hat{\mathbf{y}}$ de una red neuronal en capas (figura 3c) y las verdaderas $\mathbf{y}$. En una red con pesos $\mathbf{w}$ y entradas $\mathbf{x}$, el procedimiento consiste en: 1) una propagación hacia adelante en la que se calcula la salida de la red $\hat{\mathbf{y}}$ y se evalúa el coste $C(\hat{\mathbf{y}}, \mathbf{y})$; 2) una retropropagación del error en la que se calcula el gradiente de C usando la regla de la cadena

$$\frac{\partial C}{\partial w_{ij}} = \frac{\partial C}{\partial a_j} \cdot \frac{\partial a_j}{\partial w_{ij}}$$

donde a_j es la activación de la capa intermedia. Finalmente se actualizan los pesos: $w_{ij} \leftarrow w_{ij} - \eta \frac{\partial C}{\partial w_{ij}}$, donde η es la tasa de aprendizaje.

Aunque el trabajo de 1986 tuvo un impacto inicial significativo, el algoritmo de retropropagación no alcanzó su auge hasta finales de la década de 2000. Esto se debió a varios factores: 1) Limitaciones computacionales: el *hardware* disponible en los años 1980 y 1990 no permitía entrenar redes neuronales grandes en tiempos razonables. 2) falta de datos: no existían grandes conjuntos de datos etiquetados, que son esenciales para demostrar el potencial de las redes neuronales profundas [22]. 3) competencia de otros métodos: Las máquinas de soporte vectorial y otros enfoques dominaron el aprendizaje automático.

Los trabajos de Hopfield y Hinton, inspirados en la física, contribuyeron a revitalizar el aprendizaje automático y a reformular principios clave de las redes neuronales. Sus aportes introdujeron conceptos esenciales en la inteligencia artificial. El modelo de Hopfield desempeñó un papel fundamental en el desarrollo de la neurociencia teórica y computacional actual, atrayendo a físicos al estudio de los sistemas neuronales. También impulsó un enfoque más centrado en las propiedades emergentes de los circuitos cerebrales, abriendo nuevas perspectivas sobre el procesamiento de la información en el cerebro. Este reconocimiento a Hopfield y Hinton pone de relieve el carácter interdisciplinario de la ciencia contemporánea, donde la convergencia entre disciplinas sigue ampliando las fronteras del conocimiento.

Agradecimientos

Deseo expresar mi agradecimiento a Luis Serrano Fernández por la realización de las figuras que ilustran este comentario.

Referencias

[1] J. J. Hopfield, Neural networks and physical systems with emergent collective computational abilities, *Proceedings of the National Academy of Sciences* **79**(8), 2554 (1982).

[2] G. E. Hinton y T. J. Sejnowski, Learning and relearning in Boltzmann machines, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* **1**, 282 (1986).

[3] D. J. Amit, H. Gutfreund y H. Sompolinsky, Storing infinite numbers of patterns in a spin-glass model of neural networks, *Physical Review Letters* **55**(14), 1530 (1985).

[4] M. Mézard, G. Parisi y M. A. Virasoro, *Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications* (World Scientific, 1987).

[5] D. J. Amit, *Modeling Brain Function: The World of Attractor Neural Networks* (Cambridge University Press, 1989).

[6] S. Ramón y Cajal, Estructura de los centros nerviosos de las aves, *Revista Trimestral de Histología Normal y Patológica* **1**, 1 (1888).

[7] L. De No, Studies on the structure of the cerebral cortex, *Journal für Psychologie und Neurologie* **45**, 381 (1933).

[8] M. Khona y I. R. Fiete, Attractor and integrator networks in the brain, *Nature Reviews Neuroscience* **23**(12), 744 (2022).

[9] D. Krotov y J. J. Hopfield, Dense associative memory for pattern recognition, *Advances in Neural Information Processing Systems* **29** (NeurIPS Proceedings, 2016).

[10] M. Demircigil, J. Heusel, M. Löwe, S. Upgang y F. Vermet, On a model of associative memory with huge storage capacity, *Journal of Statistical Physics* **168**, 288 (2017).

[11] J. Hertz, A. Krogh y R. G. Palmer, *Introduction to the Theory of Neural Computation* (Addison-Wesley, 1991).

[12] E. Domany, J. L. Van Hemmen y K. Schulten, *Models of Neural Networks I* (Springer-Verlag, 1995).

[13] G. Parisi, A memory which forgets, *Journal of Physics A: Mathematical and General* **19**(10), L617 (1986).

[14] N. Parga y M. A. Virasoro, The ultrametric organization of memories in a neural network, *Journal de Physique* **47**(11), 1857 (1986).

[15] C. Van Vreeswijk y H. Sompolinsky, Chaos in neuronal networks with balanced excitatory and inhibitory activity, *Science* **274**(5293), 1724 (1996).

[16] A. Renart, J. De La Rocha, P. Bartho, L. Hollender, N. Parga, A. Reyes y K. D. Harris, The asynchronous state in cortical circuits, *Science* **327**(5965), 587 (2010).

[17] V. Mante, D. Sussillo, K.v. Shenoy y W. T. Newsome, Context-dependent computation by recurrent dynamics in prefrontal cortex, *Nature* **503**(7474), 78 (2013).

[18] F. Carnevale, V. De Lafuente, R. Romo, O. Barak y N. Parga, Dynamic control of response criterion in premotor cortex during perceptual detection under temporal uncertainty, *Neuron* **86**(4), 1067 (2015).

[19] L. Serrano-Fernández, M. Beirán y N. Parga, Emergent perceptual biases from state-space geometry in trained spiking recurrent neural networks, *Cell Reports* **43**(7), 114412 (2024).

[20] A. Vaswani, Attention Is All You Need, *Advances in Neural Information Processing Systems* (NeurIPS 30, 2017).

[21] H. Ramsauer *et al.,* Hopfield Networks Is All You Need, *Proceedings of the International Conference on Learning Representations* (ICLR, 2021).

[22] G. E. Hinton, Training products of experts by minimizing contrastive divergence, *Neural Computation* **14**(8), 1771 (2002).

[23] Y. Lecun, Y. Bengio y G. E. Hinton, Deep learning, *Nature* **521**(7553), 436 (2015).

[24] D. E. Rumelhart, G. E. Hinton y R. J. Williams, Learning representations by back-propagating errors, *Nature* **323**(6088), 533 (1986).

Néstor Parga
Depto. de Física Teórica
Universidad Autónoma
de Madrid