ÍNDICE

1
¿QUÉ ES UNA RED NEURONAL CONVOLUCIONAL?

No en mejor ocasión vamos a ser conscientes de la complejidad del proceso visual que viendo un cuadro impresionista. Un lienzo compuesto de trazos repartidos de manera local que se «sube» a la mente como una imagen con significado pleno. Aquel árbol, esta otra casa y ese barco en la lejanía aparecen ya como una escena coherente y dentro de un relato. Partes inconexas dan un todo conectado. El puntillismo hace todavía más patente el fenómeno. Un cuadro hecho mediante el uso de diminutos puntos que conforman una escena coherente.

Podríamos decir, y ahora ya hablando con más rigor, que una imagen está compuesta de propiedades primitivas. Cada punto en el lienzo, con su color y su textura, puede considerarse como una propiedad primitiva. El reto del sistema visual y sus funciones es procesar todas esas propiedades primitivas y devolver un significado holístico. Es pues un ejercicio de abstracción donde se extraen representaciones más abstractas que posibilitan una interpretación emergente a partir de propiedades primitivas. La palabra emergente es de suma importancia porque los modelos de redes neuronales convolucionales aprenden a abstraer representaciones visuales y hacer que emerjan patrones que son fruto de la concurrencia de primitivos visuales (de una pixelada de la realidad). De las propiedades primitivas obtendremos propiedades emergentes a partir de una serie de mapeados locales de manera secuencial. ¿Qué es la capacidad simbólica sino construir realidades de manera emergente?

Las redes neuronales convolucionales surgieron como una solución técnica a este reto. De hecho, aun sabiendo que existen grandes avances en inteligencia artificial casi a diario, se podría defender que es la solución más robusta hasta la fecha.[1] Aunque LeCun & Bengio (1995) concibieron las redes neuronales convolucionales para procesar también texto y series temporales, es en el campo de la

[1] Recientemente, las redes neuronales *transformer* (e. g., Devlin *et al*., 2018; Jorge-Botana, 2024; Radford *et al*., 2018; Vaswani *et al*., 2017) han ganado mucha fuerza incluso en el campo de la visión artificial o visión computacional. Se esperan grandes avances en los próximos años.

visión artificial o visión computacional donde han logrado sus mayores logros. La filosofía que hay detrás de esta arquitectura de red neuronal artificial[2] es justo la expresada: el procesamiento local de propiedades primitivas que se va integrando en diversas capas de abstracción hasta hacer un procesamiento más holístico, más elaborado. Un ejemplo clásico podría ser, justamente, los primeros prototipos de arquitecturas que utilizaron redes neuronales con capas convolucionales y retropropagación para el reconocimiento de dígitos escritos a mano (LeCun *et al.*, 1989).

La denominación «de convolución» se debe a que operan a través de convoluciones (LeCun & Bengio, 1995). Técnicamente, la convolución es una operación matemática cuyo resultado representa la magnitud en la que se superponen dos funciones o, en el contexto del procesamiento visual, dos patrones (pero no nos adelantemos más de la cuenta, ya que más tarde veremos detalladamente todo).

Para ir rompiendo el hielo, podríamos resumir el funcionamiento de las redes neuronales convolucionales en cinco pasos:

— División de la imagen en grupos adyacentes de propiedades primitivas de la entrada (*input*). A estos grupos se les puede llamar regiones (por *regions* o *patches*, su traducción al inglés), pues son grupos de píxeles que se solapan entre ellos para cubrir de manera máxima la imagen. Más tarde veremos la forma de hacerlo.

— Inferencia de patrones (o propiedades de mayor abstracción) a través de filtros actualizables según su éxito (llamados también *kernels* en inglés). Estos filtros se aplicarán a todas y cada una de las regiones en las que se divide la imagen. Este tipo de procesamiento es local (por regiones) y jerárquico (la representación aumenta en abstracción), lo que ahorra computación y distribuye la captación de patrones, cosa muy útil en la visión artificial.

— Agrupación y depuración de las señales que se van devolviendo de los distintos filtros. La información abstracta producida por los filtros es tratada para hacerla más sencilla y estable. En ciertas arquitecturas de redes neuronales convolucionales tenemos múltiples secuencias de filtrado-agrupación-depuración que se suceden.

[2] Una red neuronal artificial es un nodo o un grupo interconectado de nodos, donde un nodo es la unidad básica de procesamiento de las redes neuronales artificiales (como una neurona en el cerebro humano) y las relaciones entre nodos se expresan a partir de pesos.

— Realización de predicciones con la unificación de la información filtrada, agrupada y depurada. Cuando tenemos una representación abstracta de la imagen, estamos en disposición de realizar predicciones. Esta parte actuará como un modelo predictivo a partir del resumen de la información de la entrada. Ya no se realiza un procesamiento local y jerárquico, sino que se aplica una red neuronal clásica en la que los nodos de capas contiguas están conectados todos con todos.

Como se intuye de los puntos anteriores, la parte realmente novedosa de las redes neuronales convolucionales son los filtros mencionados en el punto 2, ya que suponen la inferencia de patrones visuales a partir de las propiedades primitivas de la entrada. Estos filtros se aplican a regiones adyacentes de la imagen para convertir esas propiedades primitivas en información sobre la presencia o ausencia de ciertos patrones más abstractos y son, sobre todo, útiles para tener éxito en la tarea que la red entera tiene que realizar. Habrá patrones mucho más útiles que otros para predecir la salida (*output*), y ahí es donde la configuración de cada filtro irá demostrando su utilidad a lo largo del entrenamiento de la red. Y lo mejor de todo, ya se ha dicho, es que esos filtros se adaptan automáticamente para detectar esos patrones sin necesidad de imponerlos a priori. Dicho esto, el lector puede imaginar que gran parte de las explicaciones en las siguientes secciones se focalizarán en estos filtros y su manera de actualizarse automáticamente.

Pero tomemos algo de altura. El listado anterior esbozaba funcionalmente las acciones más importantes que se llevan a cabo en una red neuronal convolucional genérica. No obstante, todas esas acciones tienen que ser implementadas de alguna forma. Pues bien, una red neuronal convolucional es lisa y llanamente una red neuronal artificial con una arquitectura concreta. Y como red neuronal posee conceptualmente[3] capas de nodos o unidades y conexiones ponderadas con pesos. Por tanto, todas las acciones a las que hemos aludido se producen a partir de la combinación de distintas capas especializadas a un trabajo. En general, se suelen identificar tres tipos de capas en esta arquitectura de red neuronal: capas convolucionales (*convolutional layers*), capas de agrupación (*pooling*) y capas completamente conectadas (*densely-connected layers*).

[3] Se dice conceptualmente pues a la postre, una red neuronal es básicamente un conjunto de matrices de pesos y gradientes a aplicar (i.e., tensores, concepto que veremos con mayor detalle más adelante), junto con funciones y optimizadores. No obstante, es muy útil la representación en la que los pesos están en las conexiones y los gradientes son la base para incrementarlos o decrementarlos.

Las capas convolucionales utilizan los filtros a los que hemos aludido antes (también llamados *kernels*) para extraer patrones relevantes de la información de la entrada. Cada filtro se aplica a todas y cada una de las regiones adyacentes en las que queda dividida la imagen. Por eso el procesamiento de las imágenes requiere de la segmentación en unidades de análisis más pequeñas (regiones adyacentes) que serán procesadas por las primeras capas de la red conservando sus relaciones espaciales. Esto significa que cada región tendrá sintonizado más de un filtro. Estos filtros, se aplican a modo de máscara sobre cada región y el resultado de ese filtro será una suma ponderada de las partes de esa región. Esa ponderación es justo la función del filtro. Filtrar significa dar prioridad o mayor importancia a algunas partes de la región por encima de otras. Aplicando estos filtros por todas las regiones adyacentes conseguimos detectar características locales como bordes, vértices, discontinuidades, texturas, etc., en las imágenes.

Después de las capas convolucionales, se suelen utilizar capas de agrupación para reducir la dimensionalidad de las características extraídas y hacer que la red sea más robusta ante pequeñas variaciones en la posición de los elementos en la entrada (más adelante, veremos algunos conceptos muy interesantes de esta arquitectura como, por ejemplo, la invarianza espacial). Así, las capas de agrupación combinan regiones vecinas y reducen su tamaño mediante operaciones de agrupación como el promediado o la selección del máximo valor. Finalmente, las características resultantes de las secuencias de convolución-agrupación se conectan a capas completamente conectadas, que funcionan como una red neuronal clásica, es decir, sin procesamiento local ni jerárquico, sino con conexiones densas. Como se intuye al reflexionar sobre las anteriores capas de convolución y agrupación, las capas completamente conectadas no reciben ya la representación de la imagen en base a propiedades primitivas, sino a propiedades mucho más abstractas que, por hacer una analogía útil, aunque irreal, podrían ser descritas en lenguaje natural de alguna manera similar a esta: «En esta región hay presencia de un pico y, además, se observa textura gruesa. En esta otra región no se ve presencia de pico, pero se vislumbra línea inclinada hacia la derecha». Este es el tipo de información que reciben las capas completamente conectadas y con ella se encargan de predecir la salida. Por tanto, si se quiere predecir algo, lo tiene fácil. Simplemente hay que unir toda la información procesada que proviene de la abstracción jerárquica de las regiones y ponerla de entrada en una red neuronal al uso, es decir, una en la que todos los nodos de entrada estén conectados con todos los de la siguiente capa oculta, y éstos con todos los de la siguiente capa oculta, y así hasta llegar a la capa de salida. Aquí