Unidad 2.4. Imitando al cerebro. Redes neuronales.

Métodos de Aprendizaje Profundo

> Este capítulo tiene ciertas cuestiones técnicas que no es necesario entender, se dejan para los lectores que tengan más interés pero en absoluto son un requisito para superar el curso.
>
> Simplemente hemos querido incluirlas por completitud y por dejar clalro que la Inteligencia Artificial no es alg mágico que surge de lo desconocido, sino que nace, en esencia, de un tratamiento estadístico de la información

Dentro del campo del aprendizaje automático surgen las llamadas redes neuronales que, aunque conceptualmente existen desde los años 60, se redefinieron en la década de 2010 y mostraron un rendimiento impresionante en datos de imágenes, texto y audio. Estos métodos se basan principalmente en redes neuronales artificiales, que fueron experimentadas por primera vez en la década de 1950. En aquel momento, las redes neuronales eran principalmente un tema de investigación y no se utilizaban tanto en aplicaciones prácticas. Gracias a la velocidad de las computadoras modernas y a algunas innovaciones algorítmicas, los métodos de aprendizaje profundo se utilizan actualmente de manera intensiva, especialmente en problemas de visión artificial y en gestión del lenguaje natural.

Conviene señalar que hay dos tareas fundamentales en el aprendizaje humano, a saber, la clasificación y la genreación o predicción.

Gran parte de la IA actual y de toda la revolucón que estamos viviendo tiene que ver con ambas, la generación en lo que conocemos como **IA Generativa** y la clasificaicón en tareas de visión aritifcial principalmente.

En esta sección nos ocupamos de describir la esencia de las redes neuronales que han permitido el desarrollo de las ténicas de IA genrativa más importantes.

Las Neuronas y el cerebro

El cerebro humano es un órgano increíblemente complejo y fascinante. Está compuesto por miles de millones de células llamadas neuronas, que son las unidades básicas del sistema nervioso. Estas neuronas están interconectadas en una red compleja y trabajan juntas para procesar información y controlar nuestras funciones cognitivas y corporales.

Imaginemos que el cerebro es una gran red de comunicación, donde cada neurona es como un pequeño nodo que envía y recibe mensajes. Estas neuronas se comunican entre sí a través de conexiones especializadas llamadas sinapsis. En estas sinapsis, las neuronas transmiten señales eléctricas y químicas para enviar información de un lugar a otro.

Cuando una neurona recibe una señal de otra neurona a través de sus dendritas, que son como pequeñas ramificaciones que se extienden desde la célula, se genera un impulso eléctrico. Este impulso eléctrico viaja a través del cuerpo de la neurona hacia su axón, que es como un largo cable que lleva la señal hacia las sinapsis.

En las sinapsis, la señal eléctrica se transforma en una señal química. La neurona emisora libera sustancias químicas llamadas neurotransmisores en el espacio entre las células, y estos neurotransmisores se unen a receptores en la neurona receptora, desencadenando un nuevo impulso eléctrico en esa neurona. Este proceso de señalización electroquímica se repite una y otra vez a lo largo de la red neuronal, permitiendo la comunicación y el procesamiento de información.

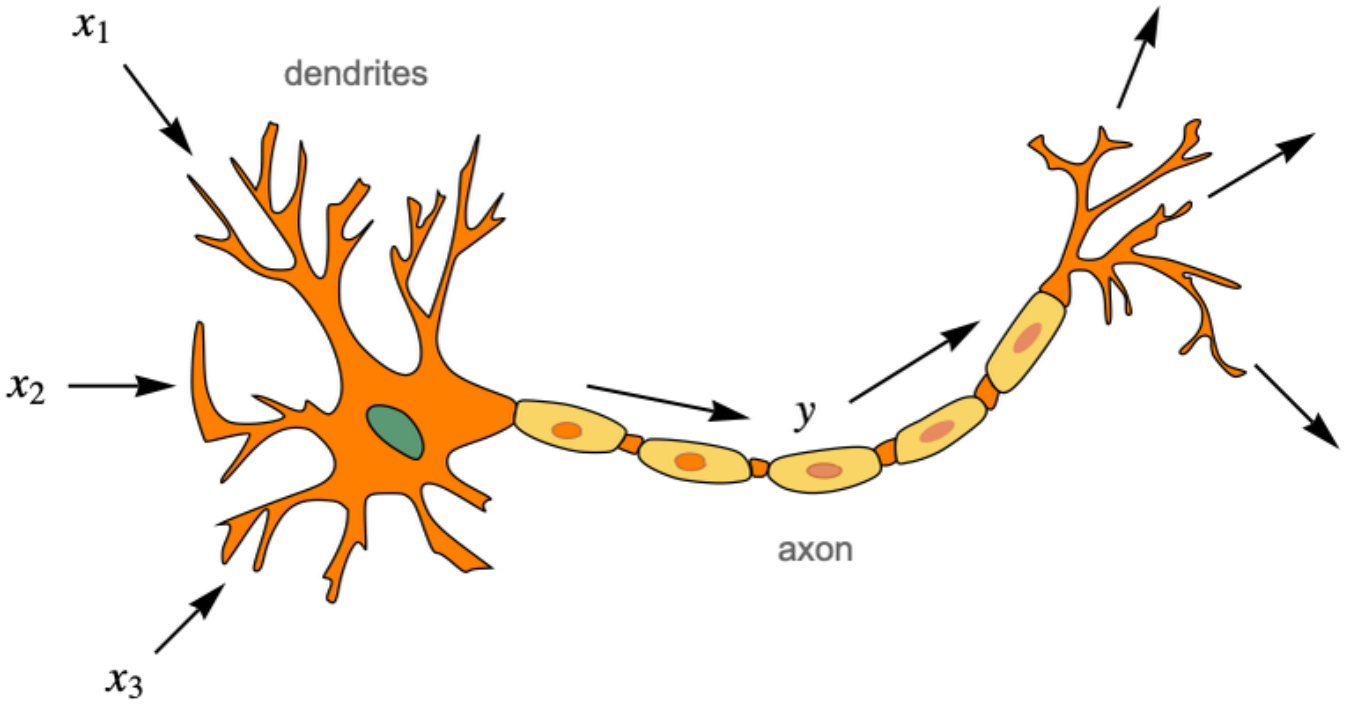
Es importante destacar que el cerebro humano no funciona de manera lineal, como una cadena de instrucciones paso a paso. En cambio, funciona de manera altamente paralela y distribuida, con múltiples neuronas trabajando simultáneamente en diferentes partes del cerebro. Esta actividad neuronal en paralelo y distribuida es lo que nos permite realizar tareas complejas como pensar, recordar, sentir emociones y realizar acciones.

El funcionamiento exacto del cerebro y cómo las neuronas procesan y almacenan información sigue siendo objeto de intensa investigación. Los neurocientíficos continúan estudiando y descubriendo nuevos aspectos sobre el funcionamiento del cerebro y cómo se relaciona con nuestras experiencias y comportamientos.

Neurona Artificial

Las redes neuronales artificiales se inspiran en lo que sabemos sobre el cerebro. En pocas palabras, el cerebro es un sistema de procesamiento de información compuesto por células llamadas neuronas, que están interconectadas en una red. Las neuronas transmiten señales

eléctricas a otras neuronas mediante estas conexiones y, juntas, son capaces de realizar cálculos que determinan nuestro comportamiento. Los seres humanos tienen alrededor de cien mil millones de neuronas en sus cerebros y aproximadamente diez mil veces más conexiones. Aquí tienes una representación clásica de lo que es una neurona biológica:

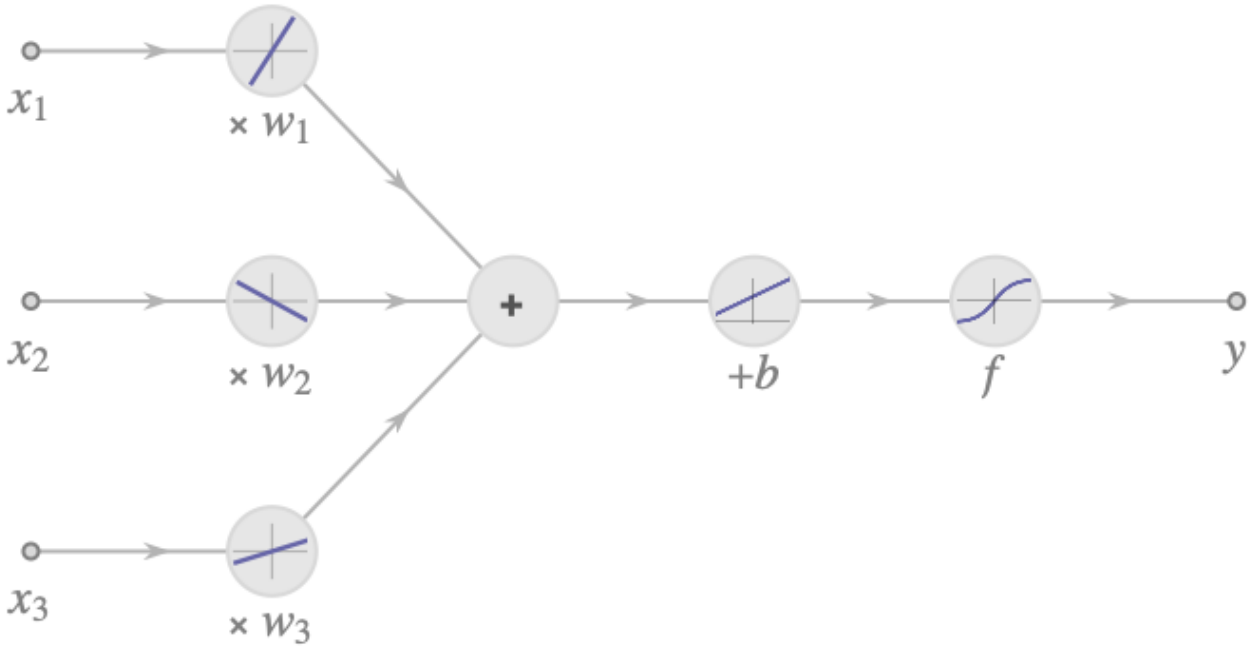


En el lado izquierdo podemos ver las dendritas, que son ramificaciones desde donde la neurona recibe sus entradas eléctricas (mostradas como x1, x2 y x3 aquí). La célula luego "calcula" una salida eléctrica (mostrada como y aquí), que viaja a lo largo del axón y se envía a muchas otras neuronas (potencialmente miles) a través de pequeñas uniones llamadas sinapsis.

Existen una gran variedad de neuronas biológicas y realizan diferentes operaciones. Tienen en común que "disparan" señales eléctricas agudas llamadas picos si se cumplen ciertas condiciones en sus entradas y estados internos. Estos cálculos analógicos son difíciles de simular y, si bien existen muchos modelos de computación de neuronas biológicas, son poco prácticos para el aprendizaje automático.

Las redes neuronales artificiales no intentan imitar exactamente las redes biológicas. En cambio, utilizan los mismos principios subyacentes al tiempo que mantienen las cosas simples y prácticas (de la misma manera que los aviones no imitan a las aves). Las redes neuronales artificiales utilizan neuronas artificiales, que son mucho más simples que sus contrapartes biológicas. Dados los valores numéricos x1, x2 y x3, la neurona artificial realiza el siguiente cálculo:

Aquí, w1, w2 y w3 son parámetros ajustables llamados pesos (los parámetros en los modelos de lenguaje), que pueden interpretarse como "fortalezas" de conexión entre neuronas. Esto podría corresponder al número de sinapsis entre dos neuronas biológicas. b es otro parámetro ajustable llamado sesgo. En las neuronas biológicas, este valor podría interpretarse como un umbral por encima del cual la neurona dispara. f es una función no lineal llamada función de activación o función de transferencia. Las neuronas biológicas también utilizan algún tipo de función de activación no lineal, ya que o bien disparan o no. Aquí tienes una ilustración del cálculo realizado por esta neurona artificial:

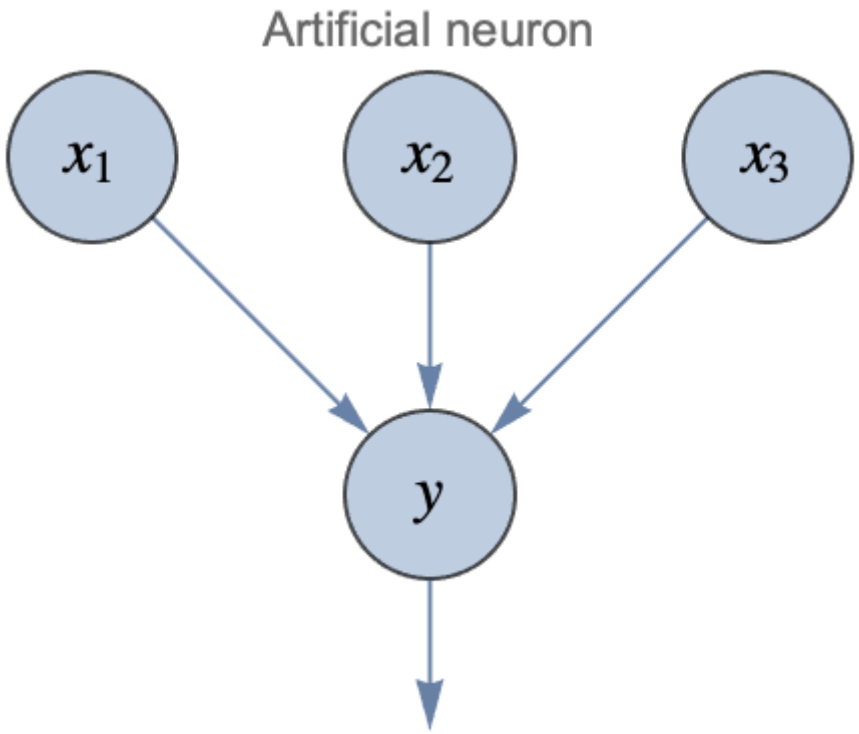


La primera parte es una combinación lineal de las características y luego se aplica una no linealidad. La presencia de esta no linealidad es importante. Permite que las redes neuronales modelen sistemas no lineales (porque la composición de funciones lineales sigue siendo una función lineal). Dado que las neuronas biológicas disparan o no disparan, resulta tentador utilizar algún tipo de función de activación escalón.

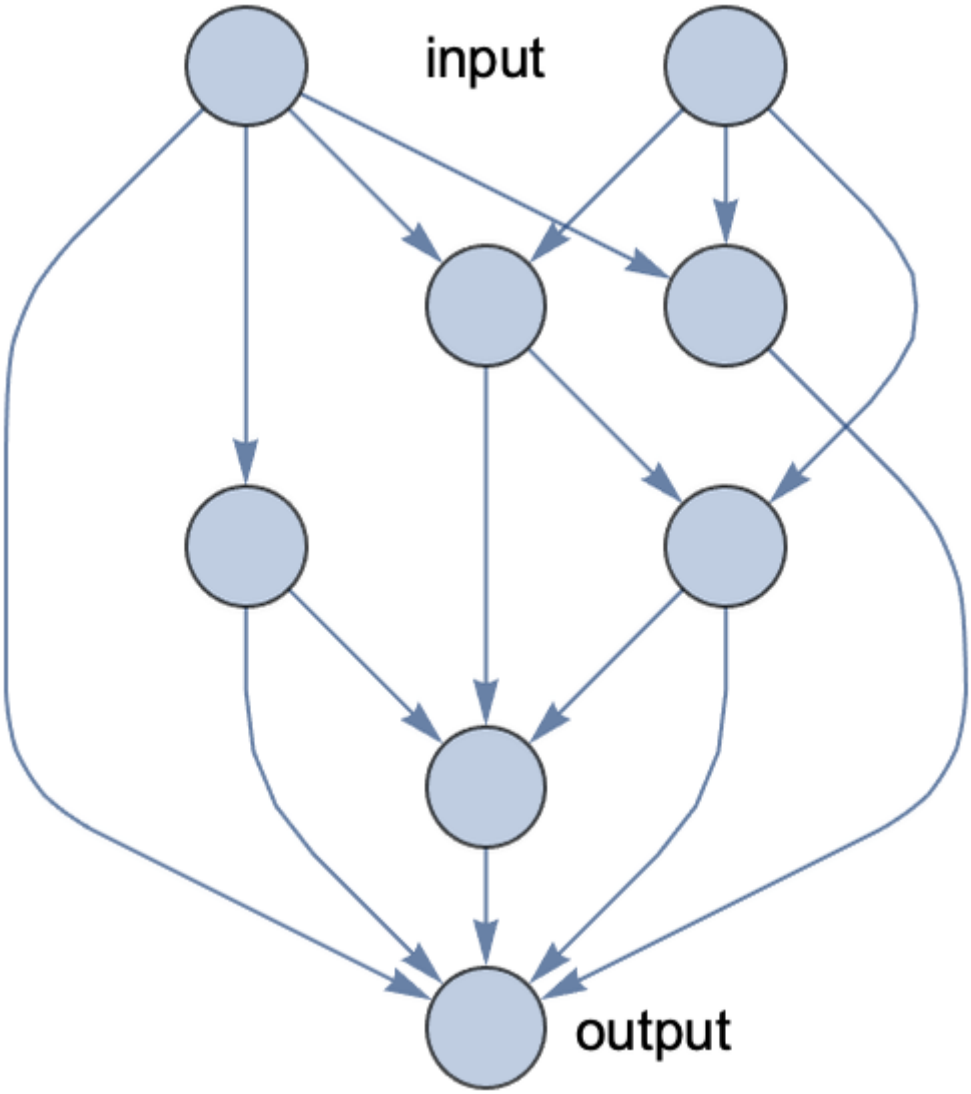
Los modelos modernos de *deep learning* tienden a alejarse de las interpretaciones biológicas. Sin embargo, sorprendentemente, siguen utilizando los mismos principios descritos aquí: combinaciones lineales de entradas seguidas de no linealidades.

Redes Neuronales

Ahora que tenemos una neurona artificial, podemos usarla para crear redes conectando muchas de ellas juntas. Cuando forman parte de una red neuronal, a menudo se les llama unidades y generalmente se representan de la siguiente manera:



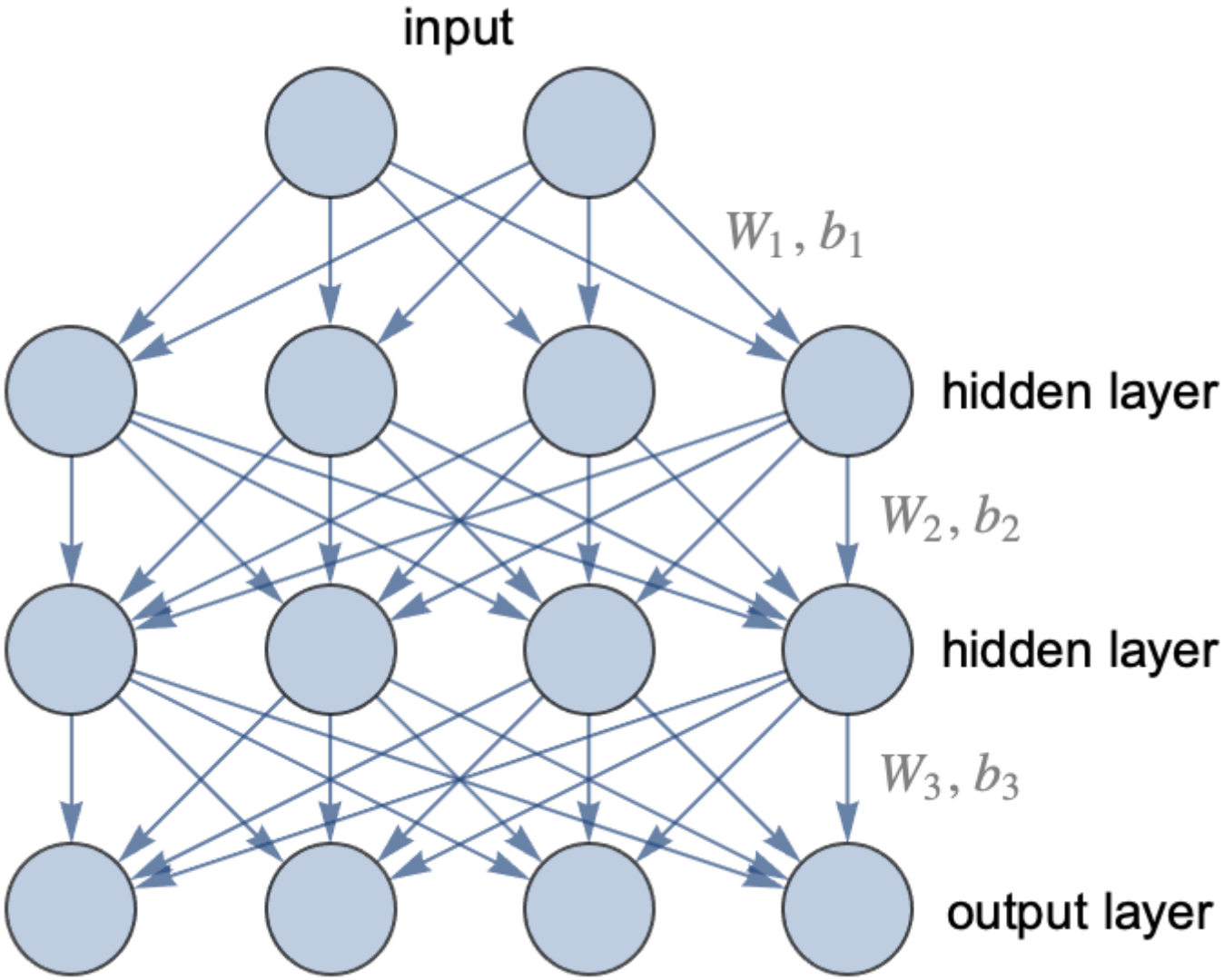
Los círculos representan valores numéricos llamados activaciones. Se da por hecho que "y" es una combinación lineal de sus entradas más algún término de sesgo y que el resultado se pasa a través de una no linealidad. Siguiendo esta convención, esto es cómo podría lucir una red neuronal artificial con conexiones aleatorias entre las neuronas:



Esta red tiene dos valores de entrada y un valor de salida. Observa que el grafo es dirigido y acíclico, por lo que podemos calcular la salida simplemente siguiendo las aristas.

Esta red es un modelo paramétrico. Hay un parámetro de peso por cada arista y un parámetro de sesgo por cada neurona. Podríamos entrenar esta red de la misma manera que cualquier otro modelo paramétrico: minimizando una función de coste calculada en algunos datos de entrenamiento. A diferencia de lo que ocurre en las redes neuronales biológicas, las aristas no se eliminan ni se añaden durante el proceso de aprendizaje; solo se modifican los parámetros numéricos (pesos y sesgos). También es posible añadir/eliminar aristas, pero es un proceso que consume tiempo, por lo que solo se realiza como un proceso separado para descubrir nuevos tipos de redes neuronales (un proceso conocido como búsqueda de arquitectura neuronal).

En la práctica, no utilizamos redes con conexiones aleatorias. En su lugar, utilizamos una arquitectura conocida. Una arquitectura neuronal no es una red exacta, sino una clase de redes neuronales que comparten estructuras similares. La arquitectura más antigua y clásica, inventada en la década de 1960, se llama perceptrón multicapa o red completamente conectada, o a veces red neuronal de propagación hacia adelante. En esta arquitectura, las neuronas se agrupan por capas, y cada neurona de una capa dada envía su salida a cada neurona de la siguiente capa (y solo a ellas). Estas capas completamente conectadas también se llaman capas lineales o capas densas. Aquí tienes un ejemplo de una red de este tipo:



Esta red tiene tres capas (la entrada no es realmente una capa): una capa de salida y dos capas llamadas capas ocultas porque son capas intermedias. Esta red toma dos valores numéricos como entrada y devuelve cuatro valores numéricos como salida. Se podría utilizar para entrenar un clasificador que tenga cuatro posibles clases, y los valores de salida serían las probabilidades de

cada clase. Para una tarea de regresión, solo tendríamos una salida (el valor predicho).

Esta arquitectura en capas permite realizar varios pasos de cálculo, al igual que lo haría un programa clásico, por lo que nos da la capacidad de realizar razonamientos. La mayoría de las redes neuronales artificiales tienen una arquitectura en capas (y las capas también están presentes en las redes neuronales biológicas). En esta ilustración, se incluyeron dos capas ocultas, pero podría haber muchas más. Una red con una capa oculta o menos se llama red superficial, mientras que una red con dos o más capas ocultas se llama red profunda, de ahí el nombre de "aprendizaje profundo". Este nombre resalta la importancia de utilizar modelos que pueden realizar varios pasos de cálculo.

En este grafo, cada flecha representa un peso. Por lo tanto, la primera capa contiene 2x4 = 8 pesos para calcular la activación de la primera capa oculta a partir de las entradas. A estos pesos, debemos agregar un parámetro de sesgo por cada salida. Estos pesos se pueden representar como una matriz.

Cada capa proporciona una serie de valores salida que se transmiten a la siguientes.

En suma el proceso general consiste en ir ajustando los pesos o parámetros de las capas de la red de forma que la salida obtenida se vaya pareciendo lo más posible a los valores reales.

El perceptrón multicapa es la arquitectura original de las redes neuronales, pero nunca logró dominar los métodos clásicos de aprendizaje automático en problemas de datos estructurados. Sin embargo, en 2017 se demostró que los perceptrones multicapa pueden competir con los métodos clásicos de aprendizaje automático en conjuntos de datos estructurados gracias a la arquitectura de auto-normalización.

Sin embargo, el uso de perceptrones multicapa sigue siendo marginal. Las redes neuronales se utilizan principalmente en datos no estructurados (imágenes, texto, sonido, etc.) gracias a arquitecturas como las redes neuronales *convolucionales*, las redes recurrentes o las redes tipo *transformer*. Estas arquitecturas todavía utilizan el concepto de capas, pero su conectividad es bastante diferente a la de los perceptrones multicapa.

Los métodos de aprendizaje profundo destacan en la resolución de problemas con datos de alta dimensionalidad, como las imágenes. En una imagen, cada píxel puede considerarse como una variable, y las interacciones entre los píxeles son cruciales para comprender el contenido general y el significado de la imagen. Un modelo superficial, como la regresión lineal, tendría dificultades para capturar estas interacciones complejas y probablemente fallaría en reconocer objetos con precisión o comprender el contenido de la imagen.

Por otro lado, los modelos de aprendizaje profundo, en particular las redes neuronales convolucionales (CNN), están diseñados específicamente para capturar y aprender patrones e

interacciones complejas dentro de la imagen. Las capas de una CNN pueden aprender representaciones jerárquicas, comenzando desde características de bajo nivel como bordes y texturas, hasta características de nivel medio como formas y objetos, y finalmente características de alto nivel que representan conceptos o categorías. Esta capacidad de extraer representaciones jerárquicas permite que los modelos de aprendizaje profundo comprendan la semántica y el contexto de la imagen, lo que les permite realizar tareas como detección de objetos, clasificación de imágenes y generación de imágenes.

La profundidad de la red permite el aprendizaje automático de características relevantes a partir de los datos, sin necesidad de ingeniería de características explícitas. Los modelos de aprendizaje profundo pueden aprender a extraer características en diferentes niveles de abstracción, lo que los hace altamente efectivos en diversas tareas de visión por computadora.

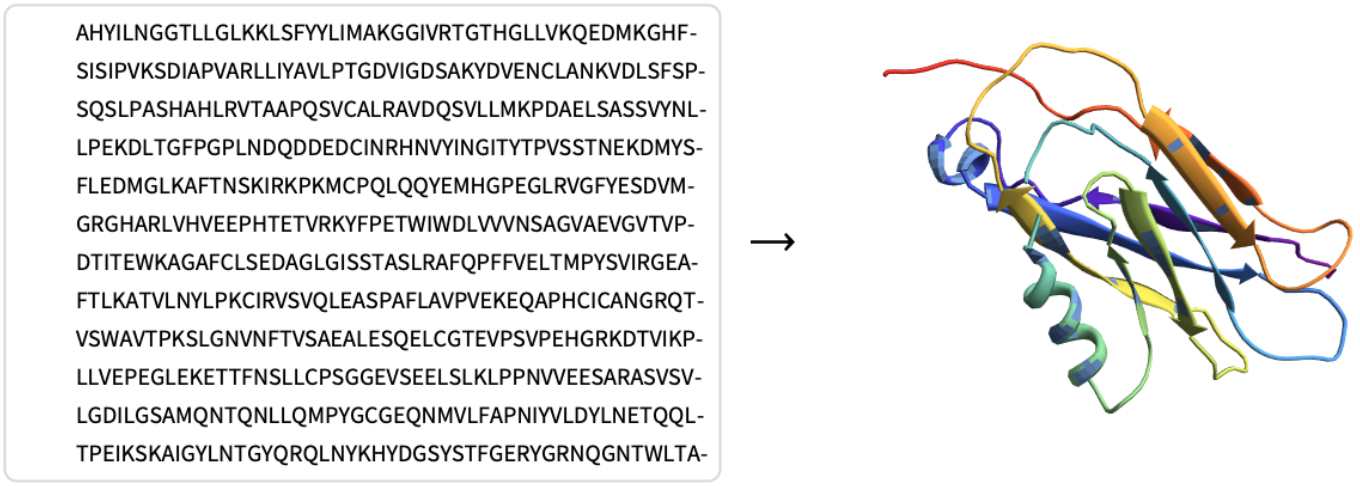
Es importante tener en cuenta que los métodos de aprendizaje profundo requieren grandes cantidades de datos etiquetados para el entrenamiento, así como recursos computacionales significativos para el entrenamiento e inferencia. Sin embargo, con los avances en hardware (*GPUs*) y la disponibilidad de conjuntos de datos a gran escala, el aprendizaje profundo se ha convertido en una herramienta poderosa en el campo de la inteligencia artificial, logrando resultados notables en diversos dominios, incluyendo visión por computadora, procesamiento del lenguaje natural y reconocimiento de voz.

https://www.youtube.com/embed/-P28LKWTzrI

Comparación del poder de una GPU fente a una CPU para pintar una imagen

Por sí mismos, los píxeles tienen muy poca información sobre lo que hay en esta imagen. No podríamos identificar este gato simplemente teniendo cada píxel "votando" sobre qué objeto es, al menos no sin que interactúen con otros píxeles. En cambio, es la interacción entre muchos píxeles, formando texturas, formas, etc., lo que proporciona información. Estos patrones no se pueden identificar con un simple cálculo. Los modelos de aprendizaje profundo pueden aprender estos patrones realizando varios cálculos simples. Al principio, detectan cosas como líneas analizando los píxeles. Luego, a partir de estas líneas, detectan formas más complejas. A partir de estas formas, detectan partes de objetos y así sucesivamente hasta que pueden identificar los principales objetos (consulte la sección Redes Convolucionales en este capítulo para una visualización de esta comprensión gradual). Este tipo de proceso de detección escalonado funciona porque la red es profunda y también porque las imágenes tienen una naturaleza algo jerárquica o compositiva. Las imágenes no son el único tipo de datos que es jerárquico/compositivo. El audio, por ejemplo, es bastante similar. El texto también lo es: los caracteres forman palabras, que luego forman oraciones, etc. El significado general surge de interacciones complejas. Las redes neuronales funcionan muy bien para todos estos tipos de datos.

Sin embargo, las redes neuronales no se limitan a imágenes, audio y texto. Por ejemplo, se utilizan para jugar juegos de mesa como el ajedrez o el Go aprendiendo a predecir si una configuración del juego es buena o no. También se utilizan para acelerar simulaciones físicas, predecir si una molécula tiene posibilidades de ser útil como medicamento o predecir cómo se pliega una proteína dada su secuencia de aminoácidos.

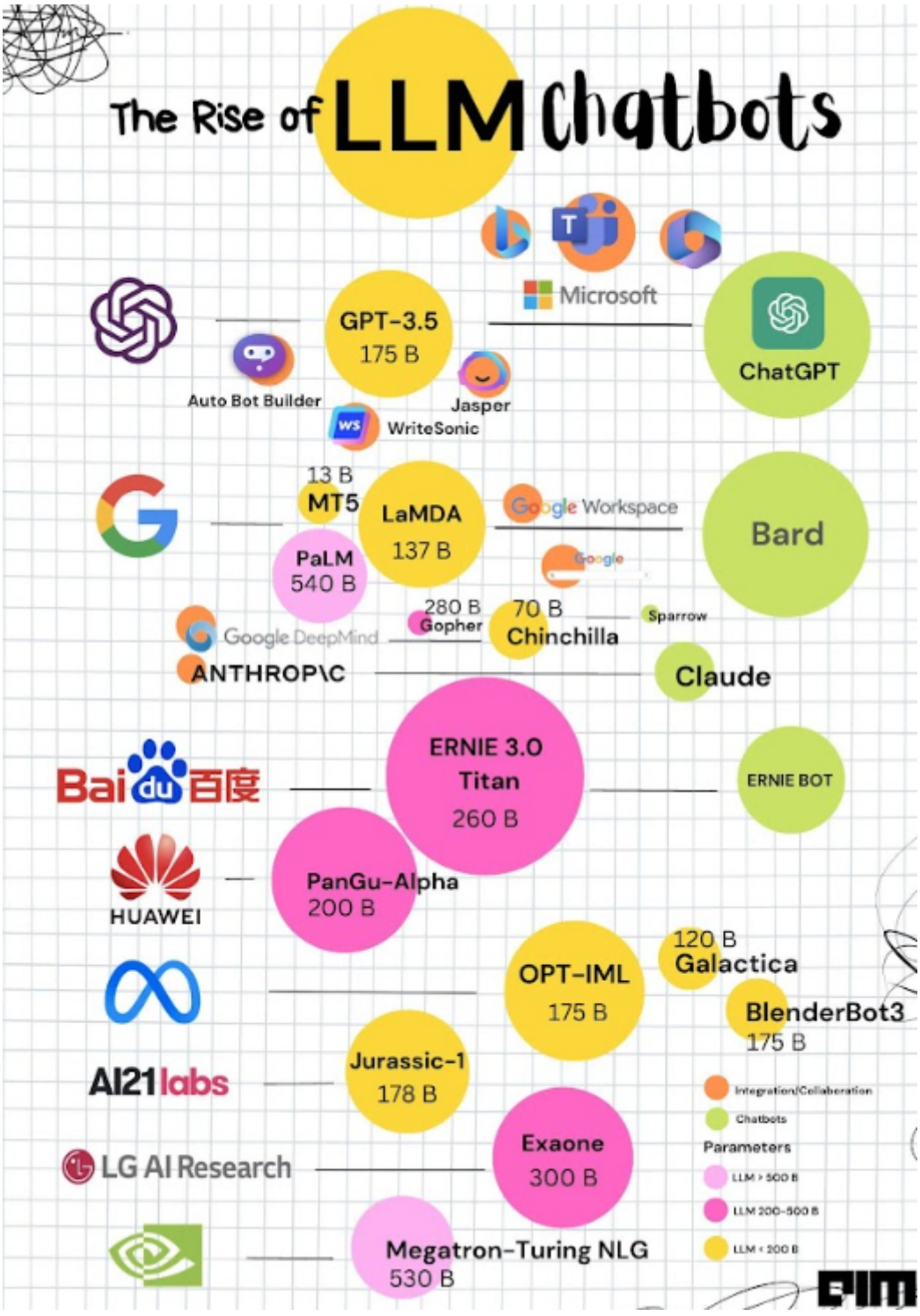


Aunque estas tareas pueden parecer diferentes de las tareas de percepción, también son tareas no estructuradas y tienen en común que los datos son de alta dimensionalidad y a menudo muestran alguna forma de composicionalidad.

Una cosa a tener en cuenta es que las redes neuronales actuales aprenden un tipo particular de programa. Por ejemplo, estas redes solo procesan matrices numéricas y no pueden manipular objetos categóricos, llamados símbolos en este contexto. Además, solo aprenden un conjunto de parámetros continuos y no aprenden construcciones de programación habituales (bucles, declaraciones condicionales, etc.). Dichos programas de red pueden ser muy buenos en algunas tareas pero no en otras. En general, las redes neuronales son bastante buenas para aprender tareas "intuitivas". Una regla general es que si un humano puede realizar una tarea en menos de un segundo, probablemente significa que una red también puede realizar esa tarea. Si le lleva más de un segundo a un humano realizar la tarea, probablemente signifique que los humanos están utilizando algún tipo de razonamiento consciente, que las redes neuronales profundas actualmente no son muy buenas modelando.

Por ejemplo, GPT-3 es un modelo de lenguaje entrenado en 2020 que tiene alrededor de 175 mil millones de parámetros. Estos tamaños son inherentes a las tareas que resuelven: reconocer imágenes requiere mucho conocimiento al igual que generar texto.

Dado que las redes neuronales profundas tienen muchos parámetros, requieren muchos ejemplos de entrenamiento. Las redes de identificación de imágenes suelen entrenarse con decenas de millones a cientos de millones de imágenes. El modelo de lenguaje GPT-3 utilizó un corpus de cientos de miles de millones de palabras. Estos conjuntos de datos contienen más datos de los que cualquier ser humano ha leído, visto u oído. Tal vez algún día descubramos redes con arquitecturas específicas que sean más eficientes en el uso de datos, pero actualmente, los conjuntos de datos grandes son esenciales para entrenar redes neuronales profundas, a menos que comencemos con una red preentrenada.



Esta infografía muestra el panorama de modelos grandes de lenguaje y sus parámetros o pesos en octubre de 2023

Actualmente los últimos modelos como chatGPT-4 trabajan con 1.75 trillones de párametros, 1000 veces más que chatGPT-3

En suma las redes usand datos numéricos de entrada para obtener una salida. En el caso de datos de tipo no numérico debemos convertirlos a números mediante diferentes técnicas que no vienne al caso en este curso.

En cualquier caso el proceso o manera en que se usan estas redes para que aprendan es similar y parte siempre de una conversión previa de los datos de entrada en números.

A continuación damos una pincelada de como funciona el proceso.

El proceso de aprendizaje

A continuación describimos el proceso de aprendizaje por pasos, pero sin entrar en detalles técnicos, solamente para entender la idea central.

Podemos dar un resumen de los estadios por lo que pasa cualquier proceso en el aprendizaje automático

1. Definir la pregunta: Es más dificil de lo que parece, y casi lo más importante, saber para que nos puede servir y que queremos resolver
2. Obtención y filtrado de datos: Necesitamos contar con datos de calidad, homogéneos y de fuentes fidedignas
3. Visualización de datos: Esto nos permite detectar errores o casos 'raros' a primera vista
4. Entrenamiento: Aquí ya es donde generamos el modelo
5. Pruebas: En esta parte usamos el grupo de datos de test para hacer las pruebas y evaluar el rendimiento o calidad del modelo
6. Analizar el feedback: Se analizan los resultados y en caso negativo se repite el proceso modificando parámetros como el número de capas y otros valores que afectan al algortitmo utilizado.
7. Usar el modelo obtenido para hacer predicciones.

A continuación describimos brevemente el proceso 4 de entrenamiento y posteriormente describiremos las distintas arquitecturas que sustentan todas las tareas de IA generativa que permiten la creación de modelos de lenguaje para el procesamiento del lenguaje natural.

1. Inicialización

Las ponderaciones y sesgos de la red se inicializan con valores pequeños y aleatorios.

2. Propagación hacia adelante (Forward Pass)

- **Entrada**: Se introduce un conjunto de entradas en la red.
- **Cálculos**: La entrada se propaga a través de las capas de la red. En cada nodo, se realiza una suma ponderada de las entradas, se le añade un sesgo y se aplica una función de activación.
- **Salida**: Se obtiene la predicción de la red para la entrada dada.

3. Cálculo de la Pérdida

- Se calcula la función de pérdida (o coste), que mide la diferencia entre la predicción de la red y la salida deseada.
- Ejemplos de funciones de pérdida incluyen el Error Cuadrático Medio para problemas de regresión y la Entropía Cruzada para problemas de clasificación.

4. Propagación hacia atrás (Backpropagation)

- **Gradiente de la Pérdida**: Se calcula el gradiente de la función de pérdida con respecto a cada uno de los pesos y sesgos en la red, usando la regla de la cadena y derivadas parciales. Esto indica cómo deberían ajustarse los pesos y sesgos para minimizar la pérdida.
- **Actualización de Pesos y Sesgos**: Se ajustan los pesos y sesgos en la dirección opuesta al gradiente para minimizar la pérdida. Esto se hace usando algoritmos de optimización como el Descenso del Gradiente o variantes más avanzadas como Adam.

5. Iteración (*backpropagation*)

Se repiten los pasos 2-4 para un número de iteraciones o épocas, utilizando diferentes conjuntos de entrada, hasta que el modelo alcance un nivel aceptable de rendimiento que se mide mediante las diferencias entre los valores reales y los valores estimados por la red.

6. Evaluación

Finalmente, se evalúa el rendimiento de la red en un conjunto de datos de prueba para asegurarse de que ha aprendido correctamente las relaciones en los datos.

Para la parte de de entrenamiento se utiliza una parte de los datos reales (*training set*) y para la evaluación y validación el resto (*validation set y test set*) .

Este proceso permite que la red neuronal multicapa ajuste sus pesos y sesgos para minimizar la función de pérdida, lo que a su vez mejora la precisión de sus predicciones en tareas de clasificación o regresión.

Este el resumen gráfico de todo el proceso:

Fase de aprendizaje

Datos entrenamiento ***Vector características*** ***Algoritmo*** ***Modelo***

Una vez entrenado ya disponemos de un modelo que nos servirá para hacer las predicciones correspondientes en el proceso siguiente:

Datos test ***Vector características*** ***Modelo*** ***Predicción***

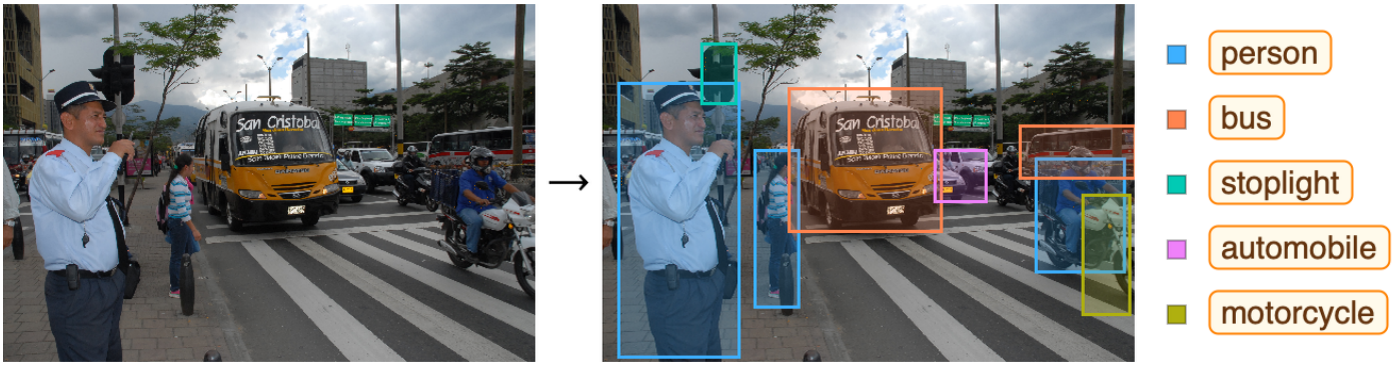
> Hemos pretendido dar una pincelada del proceso, en la sección final del módulo encontrarás referencias para estudiarlo en profundidad

Veremos ahora de forma descriptiva las principales arquitecturas neuronales y sus aplicaciones

Redes convolucionales

Las redes neuronales convolucionales (CNNs por sus siglas en inglés) son un tipo de redes neuronales que han demostrado ser muy efectivas en tareas relacionadas con la visión por computadora y el procesamiento de imágenes. Su arquitectura está inspirada en la forma en que los seres humanos procesamos la información visual.

Una aplicación típica es la detección de objetos como en este ejemplo en el que se detectan los distintos objetos representados por rectángulos.



Estructura de una CNN:

Las CNNs están compuestas por diferentes tipos de capas, cada una con una función específica:

1. Convolucional: Aplica diferentes filtros (o kernels) a la imagen de entrada para extraer características importantes como bordes, texturas, etc. Cada filtro se desliza (o convoluciona) sobre la imagen para crear un mapa de características.

2. Activación (ReLU): Introduce no linealidades en el modelo, permitiendo que la red pueda aprender patrones más complejos. La función de activación más común es la unidad lineal rectificada (ReLU).

3. Agrupación (Pooling) Reduce la dimensionalidad de los mapas de características, haciéndolos más manejables y reduciendo el riesgo de sobreajuste. El agrupamiento promedio y el agrupamiento máximo son técnicas comunes de agrupación.

4. Completamente Conectada (Fully Connected): Después de varias capas convolucionales y de agrupación, la información se aplanará en un vector y se alimentará a una o varias capas completamente conectadas, similares a las de un perceptrón multicapa.

5. Salida: La última capa produce la salida de la red, que podría ser una clasificación, una regresión, etc.

Usos Principales:
Son ampliamente utilizadas en una variedad de aplicaciones, incluyendo:

1. Reconocimiento de Imágenes: Clasificar imágenes en categorías.
2. Detección de Objetos: Localizar y clasificar objetos dentro de una imagen.
3. Segmentación Semántica: Clasificar cada píxel de una imagen en una categoría.
4. Reconocimiento Facial Identificar y verificar rostros en imágenes y vídeos.
5. Procesamiento de Lenguaje Natural: Aunque las CNNs se diseñaron originalmente para imágenes, también se han aplicado con éxito a datos de texto y voz.

Tipos de CNNs:
Existen varias arquitecturas de CNNs que se han vuelto populares y ampliamente utilizadas:

LeNet-5: Una de las primeras *CNNs*, diseñada para reconocimiento de dígitos.
AlexNet: Ganó el concurso *ImageNet* en 2012 y popularizó las CNNs en la visión por computadora.
VGGNet: Famosa por su arquitectura simple y profunda, utiliza capas convolucionales de pequeño tamaño.
GoogLeNet (*Inception*): Introduce los módulos *inception*, que son bloques de construcción que permiten a la red elegir automáticamente el tamaño del filtro en cada capa.
EfficientNet s una arquitectura de red neuronal convolucional que ha recibido una atención significativa debido a su capacidad para alcanzar un alto rendimiento en tareas de clasificación de imágenes con un número relativamente menor de parámetros en comparación con otras arquitecturas populares. Fue introducido por investigadores de Google en 2019.

Estos son solo algunos ejemplos, y la investigación en este campo está en constante evolución, con nuevas arquitecturas y técnicas que se desarrollan regularmente. Las CNNs son una herramienta poderosa en el campo del aprendizaje profundo y han revolucionado muchas aplicaciones en la visión por computadora y más allá.

Redes Recurrentes

Las Redes Neuronales Recurrentes (RNNs por sus siglas en inglés) son un tipo de red neuronal diseñado para manejar secuencias de datos, capturando la dependencia temporal entre los elementos de la secuencia. A diferencia de las redes neuronales tradicionales, las RNNs tienen conexiones recurrentes que permiten la propagación de información a través de los pasos de tiempo, haciendo posible que la salida en un momento dado dependa de los cálculos anteriores.

Características Clave de las RNNs:

1. Memoria: Las RNNs mantienen un estado interno o memoria que captura información sobre los elementos anteriores de la secuencia.

2. Conexiones Recurrentes: Cada unidad en una RNN recibe entrada no solo del dato actual en la secuencia sino también de su propio estado anterior.

3. Parámetros Compartidos: A través de todos los pasos de tiempo, una RNN utiliza el mismo conjunto de parámetros, lo que reduce significativamente la cantidad de parámetros y facilita el aprendizaje de patrones temporales.

Aplicaciones:

Las RNNs son especialmente útiles para tareas que involucran secuencias de datos, incluyendo:

1. Procesamiento de Lenguaje Natural (PLN): Traducción automática, generación de texto, reconocimiento de voz, etc.
2. Series Temporales: Predicción del mercado de valores, análisis de tendencias, etc.
3. Reconocimiento de Secuencias: Como en la escritura a mano o el reconocimiento del habla.
4. Generación de Música: Creación de melodías basadas en patrones aprendidos.

Variantes y Mejoras:

1. LSTM (Long Short-Term Memory): Una mejora con respecto a las RNNs tradicionales que pueden capturar dependencias a largo plazo en los datos gracias a su estructura especial de celda de memoria y puertas de olvido, entrada y salida.
2. GRU (Gated Recurrent Unit): Similar a LSTM pero con una estructura más simplificada, lo que la hace más eficiente en términos de computación.

Desafíos:

- Problema del Gradiente Desvaneciente y Explosivo: En las RNNs, los gradientes pueden volverse extremadamente pequeños o grandes, lo que hace que el entrenamiento sea inestable y difícil. Las LSTMs y GRUs fueron diseñadas para mitigar este problema.

Resumen

Las Redes Neuronales Recurrentes representan un avance crucial en el aprendizaje de patrones temporales y secuenciales, y han habilitado una amplia variedad de aplicaciones en el procesamiento de lenguaje natural, reconocimiento de patrones y más. Su capacidad para mantener un estado interno las hace únicas y poderosas, aunque no están exentas de desafíos, especialmente cuando se trata de secuencias muy largas. Las variantes como LSTM y GRU han demostrado ser soluciones efectivas a muchos de estos desafíos, permitiendo el aprendizaje de dependencias a largo plazo.

Redes Transformers

Las redes Transformer son un tipo de arquitectura de red neuronal introducida en el artículo “*Attention is All You Need*” por Vaswani et al. en 2017. Han revolucionado el campo del procesamiento del lenguaje natural (PLN) y se han convertido en la base para modelos como BERT, GPT, y muchos otros.

Características Clave de las Redes Transformer:

1. **Mecanismo de Atención**: En lugar de depender de recurrencias o convoluciones, las redes Transformer utilizan mecanismos de atención para ponderar la importancia de diferentes partes de la entrada en relación con cada palabra o token. La atención permite que el modelo se enfoque en las partes relevantes de la entrada para realizar la tarea y

sobre todo que tenga en cuenta el contexto en que se encuentra de modo que diferencie el significado por ejemplo de palabras iguales.

2. **Codificadores y Decodificadores**: La arquitectura Transformer original consiste en una serie de bloques de codificadores y decodificadores. Los codificadores procesan la entrada, y los decodificadores generan la salida. En tareas como la generación de texto, solo se utilizan los decodificadores.
3. **Capas de Normalización y Feedforward**: Además del mecanismo de atención, los Transformers también utilizan capas de normalización y capas feedforward para procesar los datos y realizar transformaciones.
4. **Paralelismo**: A diferencia de las RNNs, las redes Transformer permiten el procesamiento paralelo de las secuencias, lo que reduce significativamente los tiempos de entrenamiento.

Aplicaciones:

Las redes Transformer han demostrado ser extremadamente eficaces en una variedad de tareas en PLN y más allá:

1. **Traducción Automática**: Traducir texto de un idioma a otro.
2. **Generación de Texto**: Crear texto coherente y relevante dado un prompt.
3. **Comprensión del Lenguaje**: Entender el significado del texto para responder preguntas, resumir texto, etc.
4. **Reconocimiento de Voz y Generación**: Convertir voz a texto y viceversa.
5. **Clasificación de Imágenes**: Aunque fueron diseñadas inicialmente para texto, las variantes de Transformers también se han aplicado con éxito a la clasificación de imágenes y otras tareas de visión por computadora.
6. **Amplicación de texto, imagne y sonido**

> “ Veremos ejemplos de estas aplicaciones en el módulo siguiente

Desafíos y Consideraciones:

- **Requerimientos de Computación**: Los Transformers requieren una cantidad significativa de poder de cómputo y memoria, especialmente para grandes conjuntos de datos o modelos grandes.
- **Necesidad de Grandes Cantidades de Datos**: Para aprovechar al máximo su capacidad, los modelos Transformer generalmente requieren grandes cantidades de datos de entrenamiento.

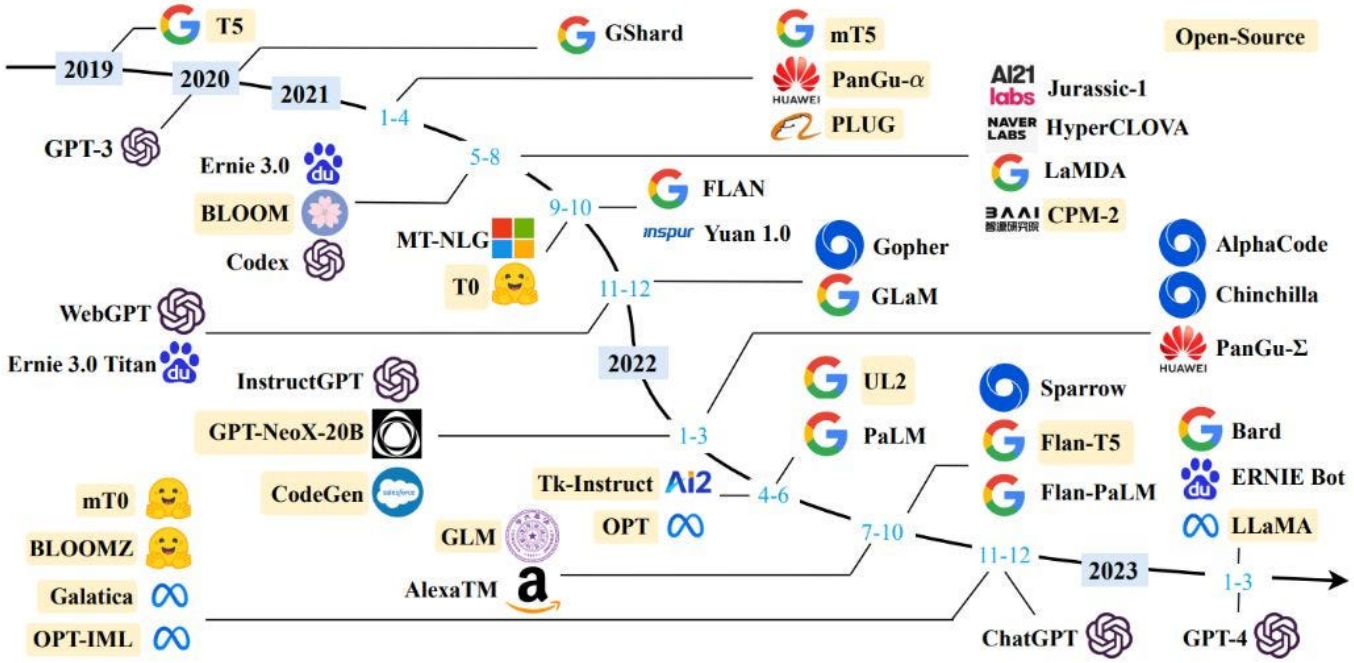
Variantes y Mejoras:

1. **BERT (Bidirectional Encoder Representations from Transformers)**: Utiliza un Transformer para aprender representaciones de palabras en base a su contexto en ambos lados (izquierda y derecha) en un texto.
2. **GPT (Generative Pre-trained Transformer)**: Un modelo de Transformer diseñado para generar texto, preentrenado en grandes cantidades de texto y afinado para tareas específicas.

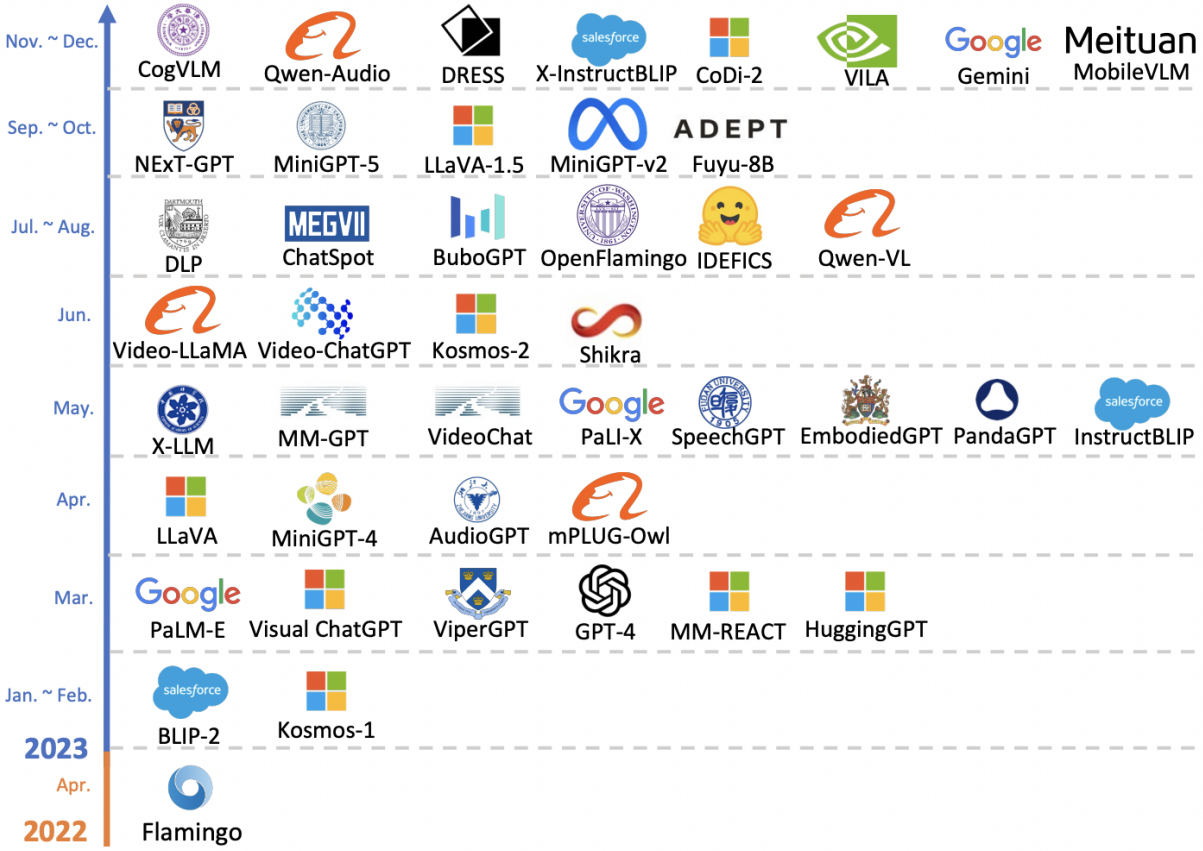
Resumen:

Las redes Transformer han establecido un nuevo estándar en PLN, ofreciendo un rendimiento excepcional en una amplia gama de tareas. Su diseño permite el procesamiento paralelo y el aprendizaje de relaciones complejas en los datos, aunque esto viene con un costo computacional significativo. Las mejoras y variantes continúan evolucionando, ampliando las capacidades y aplicaciones de esta poderosa arquitectura.

Para acabar esta sección dejamos un esquema de la evolución del ecosistema de modelos de tipo *transformer* generados y entrenados hasta el presente.



Fuente: https://www.wolfram.com/language/introduction-machine-learning/deep-learning-methods/



Este es el panorama de chatbots de empresas a finales de 2023

Glosario de términos

Añadimos finalmente un glosario de términos de uso común en el ecosistema de la IA

Término	Descripción
Token	Una unidad básica de procesamiento de texto, como una palabra, un número o un signo de puntuación.
Contexto	El conjunto de tokens o palabras que rodean a otro token específico, importante para determinar su significado.
Modelo de lenguaje	Un sistema de inteligencia artificial diseñado para entender, generar y manipular lenguaje humano.
Aprendizaje profundo	Una técnica de aprendizaje automático que enseña a los ordenadores a aprender realizando tareas.
Red neuronal	Un modelo computacional inspirado en la forma en que funcionan los cerebros humanos.

Transformer	Un tipo de arquitectura de red neuronal especializada en procesamiento de lenguaje.
Entrenamiento	El proceso de enseñar a un modelo de lenguaje a entender y generar texto a través de ejemplos.
Generación de texto	La habilidad de un modelo de lenguaje para crear texto nuevo y coherente.
Comprensión del lenguaje	La capacidad de un modelo de lenguaje para entender el significado y la intención detrás del texto.
API (Interfaz de Programación de Aplicaciones)	Una interfaz que permite la interacción con un modelo de lenguaje a través de programas.
Aprendizaje supervisado	Un tipo de aprendizaje automático donde el modelo se entrena con datos etiquetados.
Parámetro de entrenamiento	Valores ajustables en un modelo de aprendizaje automático que se optimizan durante el entrenamiento.
Dimensionalidad	Se refiere al número de atributos o características que tienen los datos de entrada en un modelo.
Aprendizaje no supervisado	Un tipo de aprendizaje automático que utiliza datos no etiquetados para encontrar patrones.
Regularización	Técnicas usadas para reducir el sobreajuste en modelos de aprendizaje automático.
Activación	Función matemática usada en redes neuronales para determinar la salida de un nodo.
Aprendizaje por refuerzo	Un tipo de aprendizaje automático donde un agente aprende a tomar decisiones a través de recompensas.
Sobreajuste (Overfitting)	Cuando un modelo de aprendizaje automático se ajusta demasiado a los datos de entrenamiento.
Subajuste (Underfitting)	Cuando un modelo de aprendizaje automático no puede capturar la estructura subyacente de los datos.

Revision #29
Created 2023-07-03 13:44:55 CEST by Luis Hueso
Updated 2024-03-16 12:41:39 CET by Luis Hueso