¿Qué son los transformadores en la inteligencia artificial?

Los transformadores son un tipo de arquitectura de red neuronal que transforma o cambia una secuencia de entrada en una secuencia de salida. Para ello, aprenden el contexto y rastrean las relaciones entre los componentes de la secuencia. Por ejemplo, considere esta secuencia de entrada: "¿De qué color es el cielo?". El modelo transformador usa una representación matemática interna que identifica la relevancia y la relación entre las palabras color, cielo y azul. Usa esa información para generar el resultado: "El cielo es azul".

Las organizaciones usan modelos de transformadores para todo tipo de conversiones de secuencias, desde el reconocimiento de voz hasta la traducción automática y el análisis de secuencias de proteínas.

Más información sobre las redes neuronales

Más información sobre la inteligencia artificial (IA)

¿Por qué son importantes los transformadores?

Los modelos tempranos de aprendizaje profundo que se centraron ampliamente en las tareas de procesamiento de lenguaje natural (NLP) tenían como fin lograr que las computadoras comprendan y respondan al lenguaje humano natural. Adivinaron la palabra siguiente en una secuencia basada en la palabra anterior.

Para entenderlo mejor, piense en la función de autocompletar de su smartphone. Hace sugerencias en función de la frecuencia de los pares de palabras que escribe. Por ejemplo, si escribe con frecuencia "Estoy bien", el teléfono sugiere automáticamente la palabra *"bien"* después de escribir *"estoy"*.

Los primeros modelos de machine learning (ML) aplicaban una tecnología similar a una escala más amplia. Trazaban la frecuencia de relación entre diferentes pares de palabras o grupos de palabras en su conjunto de datos de entrenamiento e intentaban adivinar la siguiente palabra. Sin embargo, la tecnología primitiva no podía retener el contexto más allá de una determinada longitud de entrada. Por ejemplo, uno de los primeros modelos de ML no podía generar un párrafo significativo porque no podía retener el contexto entre la primera y la última oración de un párrafo. Para generar

un resultado como “Soy de Italia. Me gusta andar a caballo. Hablo italiano”, el modelo debe recordar la conexión entre Italia e italiano, algo que las primeras redes neuronales simplemente no podían hacer.

Los modelos de transformadores cambiaron radicalmente las tecnologías de NLP al permitir que los modelos administraran dependencias de largo alcance en el texto. Los siguientes son más beneficios de los transformadores.

Habilitar modelos a gran escala

Los transformadores procesan secuencias largas en su totalidad con cálculo paralelo, lo que reduce significativamente tanto el tiempo de entrenamiento como el de procesamiento. Esto ha permitido el entrenamiento de modelos de lenguaje de gran tamaño (LLM), como GPT y BERT, que pueden aprender representaciones lingüísticas complejas. Tienen miles de millones de parámetros que capturan una amplia gama de conocimientos y lenguaje humanos, y están impulsando la investigación hacia sistemas de inteligencia artificial más generalizables.

Más información sobre los modelos de lenguaje de gran tamaño

Más información sobre GPT

Facilitar una personalización más rápida

Con los modelos de transformadores, puede utilizar técnicas como el aprendizaje por transferencia y la generación aumentada de recuperación (RAG). Estas técnicas permiten la personalización de los modelos existentes para aplicaciones específicas de la organización del sector. Los modelos pueden entrenarse previamente en conjuntos de datos grandes y, a continuación, ajustarse con precisión en conjuntos de datos más pequeños y específicos para tareas específicas. Este enfoque ha democratizado el uso de modelos sofisticados y ha eliminado las limitaciones de recursos en el entrenamiento de modelos grandes desde cero. Los modelos pueden funcionar bien en varios dominios y tareas para diversos casos de uso.

Facilite los sistemas de IA multimodales

Con los transformadores, puede usar la IA para tareas que combinan conjuntos de datos complejos. Por ejemplo, modelos como el DALL-E muestran que los transformadores pueden generar imágenes a partir de descripciones textuales, combinando capacidades de NLP y visión artificial. Con los transformadores, puede crear aplicaciones de inteligencia artificial que integren diferentes tipos de información e imiten más de cerca la comprensión y la creatividad humanas.

Más información sobre la visión artificial

Investigación de IA e innovación industrial

Los transformadores han creado una nueva generación de tecnologías e investigaciones de IA, ampliando los límites de lo que es posible en el ML. Su éxito ha inspirado nuevas arquitecturas y aplicaciones que resuelven problemas innovadores. Han permitido que las máquinas entiendan y generen el lenguaje humano, lo que ha dado como resultado aplicaciones que mejoran la experiencia del cliente y crean nuevas oportunidades de negocio.

¿Cuáles son los casos de uso de los transformadores?

Usted puede entrenar modelos de transformadores de gran tamaño con cualquier dato secuencial, como lenguajes humanos, composiciones musicales, lenguajes de programación y más. A continuación, se presentan algunos ejemplos de casos de uso.

Procesamiento del lenguaje natural

Los transformadores les permiten a las máquinas entender, interpretar y generar el lenguaje humano de una manera más precisa que nunca. Pueden resumir documentos de gran tamaño y generar texto coherente y contextualmente relevante para todo tipo de casos de uso. Los asistentes virtuales como Alexa utilizan la tecnología de los transformadores para entender y responder a los comandos de voz.

Traducción automática

Las aplicaciones de traducción utilizan transformadores para proporcionar traducciones precisas y en tiempo real entre idiomas. Los transformadores han mejorado significativamente la fluidez y precisión de las traducciones en comparación con las tecnologías anteriores.

Más información sobre la traducción automática

Análisis de secuencias de ADN

Al tratar los segmentos del ADN como una secuencia similar al lenguaje, los transformadores pueden predecir los efectos de las mutaciones genéticas, comprender los patrones genéticos y ayudar a identificar las regiones del ADN que son responsables de ciertas enfermedades. Esta capacidad es fundamental para la medicina personalizada, en la que comprender la composición genética de una persona puede conducir a tratamientos más eficaces.

Análisis de la estructura de las proteínas

Los modelos de transformadores pueden procesar datos secuenciales, lo que los hace muy adecuados para modelar las cadenas largas de aminoácidos que se pliegan en estructuras proteicas complejas. Comprender las estructuras de las proteínas es fundamental para el descubrimiento de fármacos y la comprensión de los procesos biológicos. También puede utilizar transformadores en aplicaciones que predicen la estructura 3D de las proteínas en función de sus secuencias de aminoácidos.

¿Cómo funcionan los transformadores?

Las redes neuronales han sido el método líder en varias tareas de la IA, como el reconocimiento de imágenes y la NLP, desde principios de la década del 2000. Consisten en capas de nodos informáticos interconectados, o *neuronas*, que imitan al cerebro humano y trabajan juntos para resolver problemas complejos.

Las redes neuronales tradicionales que se ocupan de secuencias de datos suelen utilizar un patrón de arquitectura de codificador/decodificador. El codificador lee y procesa toda la secuencia de datos de entrada, como una oración en inglés, y la transforma en una representación matemática compacta. Esta representación es un resumen que captura la esencia de la entrada. Luego, el decodificador toma este resumen y, paso a paso, genera la secuencia de salida, que podría ser la misma oración traducida al francés.

Este proceso ocurre de forma secuencial, lo que significa que tiene que procesar cada palabra o parte de los datos una tras otra. El proceso es lento y puede perder algunos detalles más finos en largas distancias.

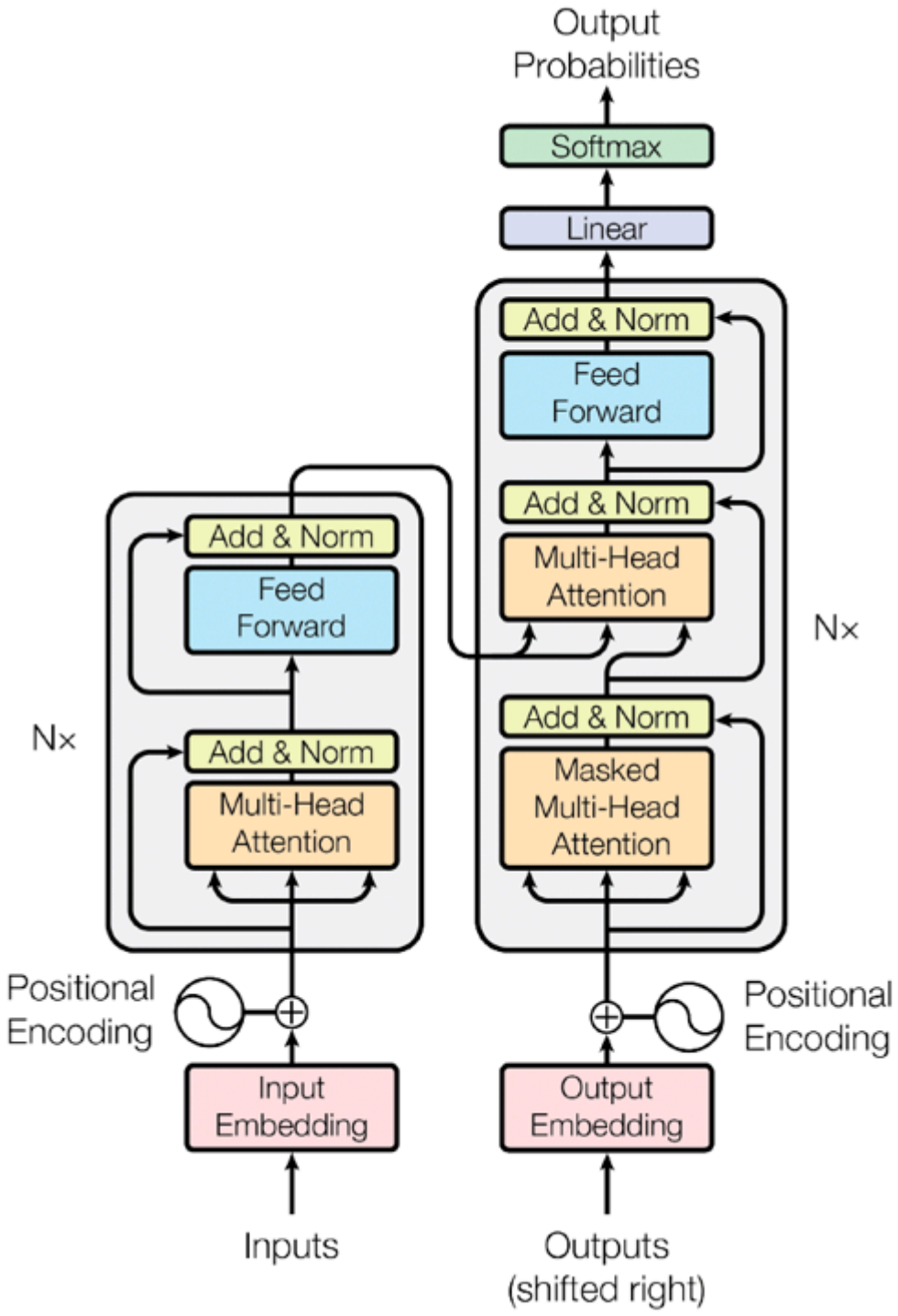
Mecanismo de autoatención

Los modelos de transformadores modifican este proceso al incorporar algo llamado *mecanismo de autoatención*. En lugar de procesar los datos en orden, el mecanismo le permite al modelo observar diferentes partes de la secuencia a la vez y determinar qué partes son las más importantes.

Imagine que está en una habitación concurrida e intenta escuchar a alguien hablar. El cerebro se centra automáticamente en su voz mientras desconecta los ruidos menos importantes. La autoatención permite al modelo hacer algo similar: presta más atención a los fragmentos de información relevantes y los combina para hacer mejores predicciones de resultados. Este mecanismo hace que los transformadores sean más eficientes, lo que les permite entrenarse en conjuntos de datos más grandes. También es más eficaz, especialmente cuando se trata de textos largos en los que el contexto lejano puede influir en el significado de lo que viene después.

¿Cuáles son los componentes de la arquitectura de los transformadores?

La arquitectura de la red neuronal del transformador tiene varias capas de software que trabajan juntas para generar el resultado final. La siguiente imagen muestra los componentes de la arquitectura de transformación, tal como se explica en el resto de esta sección.



Incrustaciones de entrada

Esta etapa convierte la secuencia de entrada en el dominio matemático que entienden los algoritmos de software. Al principio, la secuencia de entrada se divide en una serie de tokens o

componentes de secuencia individuales. Por ejemplo, si la entrada es una oración, los tokens son palabras. La incrustación transforma entonces la secuencia del token en una secuencia vectorial matemática. Los vectores contienen información semántica y sintáctica, representada como números, y sus atributos se aprenden durante el proceso de entrenamiento.

Puede visualizar los vectores como una serie de coordenadas en un espacio *n*-dimensional. Como ejemplo simple, piense en una gráfica bidimensional, donde *x* representa el valor alfanumérico de la primera letra de la palabra e *y* representa sus categorías. La palabra *banana* tiene el valor (2,2) porque comienza con la letra *b* y está en la categoría *fruta*. La palabra *mango* tiene el valor (13,2) porque comienza con la letra *m* y también está en la categoría *fruta.* De esta forma, el vector (*x, y*) le dice a la red neuronal que las palabras *banana* y *mango* pertenecen a la misma categoría.

Ahora imagine un espacio *n*-dimensional con miles de atributos sobre la gramática, el significado y el uso de cualquier palabra en oraciones asignadas a una serie de números. El software puede usar los números para calcular las relaciones entre las palabras en términos matemáticos y comprender el modelo del lenguaje humano. Las incrustaciones permiten representar los tokens discretos como vectores continuos que el modelo puede procesar y de los que puede aprender.

Codificación posicional

La codificación posicional es un componente crucial en la arquitectura del transformador porque el modelo en sí no procesa inherentemente los datos secuenciales en orden. El transformador necesita una forma de considerar el orden de las fichas en la secuencia de entrada. La codificación posicional agrega información a la incrustación de cada token para indicar su posición en la secuencia. Esto se hace a menudo mediante el uso de un conjunto de funciones que generan una señal posicional única que se agrega a la incrustación de cada token. Con la codificación posicional, el modelo puede preservar el orden de los símbolos y comprender el contexto de la secuencia.

Bloque transformador

Un modelo de transformador típico tiene varios bloques de transformadores apilados juntos. Cada bloque transformador tiene dos componentes principales: un mecanismo de autoatención con múltiples cabezales y una red neuronal de retroalimentación por posición. El mecanismo de

autoatención permite al modelo sopesar la importancia de las diferentes fichas dentro de la secuencia. Al hacer predicciones, se centra en las partes relevantes de la entrada.

Por ejemplo, piense en las frases "*Speak no lies (No diga mentiras)*" y "*He lies down (Se acuesta)".* En ambas oraciones, el significado de la palabra *lies* no se puede entender sin mirar las palabras que están al lado. Las palabras *speak* y *down* son fundamentales para entender el significado correcto. La autoatención permite agrupar los elementos relevantes para el contexto.

La capa de alimentación directa tiene componentes adicionales que ayudan al modelo del transformador a entrenarse y funcionar de manera más eficiente. Por ejemplo, cada bloque transformador incluye lo siguiente:

- Conexiones en torno a los dos componentes principales que actúan como accesos directos. Permiten el flujo de información de una parte de la red a otra, omitiendo ciertas operaciones intermedias.
- La normalización de capas mantiene los números (específicamente las salidas de las diferentes capas de la red) dentro de un rango determinado para que el modelo se capacite sin problemas.
- La transformación lineal funciona para que el modelo ajuste los valores para realizar mejor la tarea en la que se está entrenando, como el resumen del documento en lugar de la traducción.

Bloques lineales y softmax

En última instancia, el modelo necesita hacer una predicción concreta, como elegir la siguiente palabra de una secuencia. Aquí es donde entra en juego el bloque lineal. Es otra capa totalmente conectada, también conocida como capa densa, antes de la etapa final. Realiza un trazado lineal aprendido desde el espacio vectorial hasta el dominio de entrada original. Esta capa crucial es donde la parte del modelo para la toma de decisiones toma las complejas representaciones internas y las convierte de nuevo en predicciones específicas que puede interpretar y utilizar. El resultado de esta capa es un conjunto de puntuaciones (a menudo denominadas logits) para cada token posible.

La función softmax es la etapa final que toma las puntuaciones logit y las normaliza en una distribución de probabilidad. Cada elemento de la salida de softmax representa la confianza del modelo en una clase o token en particular.

¿En qué se diferencian los transformadores de otras arquitecturas de redes neuronales?

Las redes neuronales recurrentes (RNN) y las redes neuronales convolucionales (CNN) son otras redes neuronales que se utilizan con frecuencia en tareas de machine learning y aprendizaje profundo. A continuación, se exploran sus relaciones con los transformadores.

Transformadores en comparación con las RNN

Tanto los modelos de transformadores como las RNN son arquitecturas que se utilizan para procesar datos secuenciales.

Las RNN procesan las secuencias de datos un elemento a la vez en iteraciones cíclicas. El proceso comienza cuando la capa de entrada recibe el primer elemento de la secuencia. Luego, la información se pasa a una capa oculta, que procesa la entrada y pasa la salida al siguiente paso de tiempo. Esta salida, combinada con el siguiente elemento de la secuencia, se retroalimenta a la capa oculta. Este ciclo se repite para cada elemento de la secuencia, y el RNN mantiene un vector de estado oculto que se actualiza en cada paso de tiempo. Este proceso le permite eficazmente a la RNN recordar información de entradas pasadas.

Por el contrario, los transformadores procesan secuencias completas simultáneamente. Esta paralelización permite tiempos de entrenamiento mucho más rápidos y la capacidad de manejar secuencias mucho más largas que las RNN. El mecanismo de autoatención de los transformadores también permite que el modelo analice toda la secuencia de datos simultáneamente. Esto elimina la necesidad de recurrencia o de vectores ocultos. En cambio, la codificación posicional mantiene la información sobre la posición de cada elemento en la secuencia.

Los transformadores han superado en gran medida a las RNN en muchas aplicaciones, especialmente en tareas de NLP, porque pueden gestionar las dependencias de largo alcance de forma más eficaz. También tienen mayor escalabilidad y eficiencia que las RNN. Las RNN siguen siendo útiles en ciertos contextos, especialmente cuando el tamaño del modelo y la eficiencia computacional son más importantes que la captura de interacciones a larga distancia.

Transformadores en comparación con las CNN

Las CNN están diseñadas para datos en forma de cuadrícula, como imágenes, donde las jerarquías espaciales y la localidad son clave. Utilizan capas convolucionales para aplicar filtros en una entrada, capturando patrones locales a través de estas vistas filtradas. Por ejemplo, en el procesamiento de imágenes, las capas iniciales pueden detectar bordes o texturas, y las capas más profundas reconocen estructuras más complejas, como formas u objetos.

Los transformadores se diseñaron principalmente para manejar datos secuenciales y no podían procesar imágenes. Los modelos de transformadores de visión ahora procesan imágenes convirtiéndolas en un formato secuencial. Sin embargo, las CNN siguen siendo una opción muy eficaz y eficiente para muchas aplicaciones prácticas de visión artificial.

¿Cuáles son los diferentes tipos de modelos de transformadores?

Los transformadores se han convertido en una familia diversa de arquitecturas. Los siguientes son algunos tipos de modelos de transformadores.

Transformadores bidireccionales

Las representaciones de codificadores bidireccionales de los modelos de transformadores (BERT) modifican la arquitectura base para procesar las palabras en relación con todas las demás palabras de una oración y no de forma aislada. Técnicamente, emplea un mecanismo llamado modelo de lenguaje enmascarado bidireccional (MLM). Durante el preentrenamiento, las BERT ocultan aleatoriamente un porcentaje de los tokens de entrada y predicen estos tokens ocultos basándose en su contexto. El aspecto bidireccional proviene del hecho de que las BERT tienen en cuenta las secuencias de tokens de izquierda a derecha y de derecha a izquierda en ambas capas para una mayor comprensión.

Transformadores generativos preentrenados

Los modelos de transformadores generativos preentrenados (GPT) utilizan decodificadores de transformadores apilados que se entrenan previamente en un gran corpus de texto mediante el uso de objetivos de modelado del lenguaje. Son autorregresivos, lo que significa que retroceden o predicen el siguiente valor de una secuencia en función de todos los valores anteriores. Al utilizar más de 175 000 millones de parámetros, los modelos de GPT pueden generar secuencias de texto

que se ajustan según el estilo y el tono. Los modelos de GPT han impulsado la investigación en IA para lograr la inteligencia artificial general. Esto significa que las organizaciones pueden alcanzar nuevos niveles de productividad y, al mismo tiempo, reinventar sus aplicaciones y experiencias de los clientes.

Transformadores bidireccionales y autorregresivos

Un transformador bidireccional y autorregresivo (BART) es un tipo de modelo de transformador que combina propiedades bidireccionales y autorregresivas. Es como una mezcla del codificador bidireccional de BERT y el decodificador autorregresivo de GPT. Lee toda la secuencia de entrada a la vez y es bidireccional como BERT. Sin embargo, genera la secuencia de salida de un token a la vez, condicionada a los tokens generados previamente y a la entrada proporcionada por el codificador.

Transformadores para tareas multimodales

Los modelos de transformadores multimodales, como ViLBERT y VisualBERT, están diseñados para administrar varios tipos de datos de entrada, normalmente texto e imágenes. Amplían la arquitectura del transformador mediante el uso de redes de doble flujo que procesan entradas visuales y textuales por separado antes de fusionar la información. Este diseño permite que el modelo aprenda las representaciones intermodales. Por ejemplo, ViLBERT usa capas transformadoras coatencionales para permitir que los flujos separados interactúen. Es crucial para situaciones en las que es clave comprender la relación entre el texto y las imágenes, como las tareas visuales de respuesta a preguntas.

Transformadores de visión

Los transformadores de visión (ViT) reutilizan la arquitectura del transformador para tareas de clasificación de imágenes. En lugar de procesar una imagen como una cuadrícula de píxeles, ven los datos de la imagen como una secuencia de parches de tamaño fijo, de forma similar a como se tratan las palabras en una oración. Cada parche se aplana, se incrusta linealmente y luego se procesa secuencialmente mediante el codificador de transformador estándar. Se agregan incrustaciones posicionales para mantener la información espacial. Este uso de la autoatención global permite al modelo capturar las relaciones entre cualquier par de parches, independientemente de su posición.