COLOMBIA POTENCIA DE LA VIDA

TIC

Modelos de Lenguaje

TALENTO TECH | AZ PROYECTOS EDUCATIVOS | UTP Universidad Tecnológica de Pereira

Modelos de Lenguaje

El modelado de lenguaje en el procesamiento del lenguaje natural es una técnica fundamental que implica la construcción de modelos estadísticos o computacionales para capturar y predecir la estructura y el comportamiento del lenguaje humano. Consiste en desarrollar modelos que puedan asignar probabilidades a secuencias de palabras en un determinado idioma. Estos modelos pueden ser utilizados en una variedad de aplicaciones en NLP, como reconocimiento de voz, traducción automática, generación de texto, corrección gramatical y más.

1. Modelos de N-grama

Los modelos de N-grama son una técnica fundamental en el procesamiento del lenguaje natural (NLP) que se utiliza para modelar la probabilidad de una secuencia de palabras en un texto. En esencia, un N-grama es una secuencia contigua de N palabras de un texto. Los modelos de N-grama son ampliamente utilizados en tareas de modelado de lenguaje, clasificación de texto, reconocimiento de patrones y más.

Un N-grama es una secuencia de N palabras que ocurren juntas en un texto.

Por ejemplo, en la frase "El gato está durmiendo en el sofá", los 2-gramas (o bigramas) serían: "El gato", "gato está", "está durmiendo", "durmiendo en", "en el", "el sofá".

Los N-gramas capturan la estructura local del lenguaje, es decir, la probabilidad de que una palabra aparezca dada su historia reciente de palabras anteriores.

Uni-Gram	This	Is	Big	Data	AI	Book
Bi-Gram	This is	Is Big	Big Data	Data AI	AI Book	
Tri-Gram	This is Big	Is Big Data	Big Data AI	Data AI Book		

Modelado de la Probabilidad

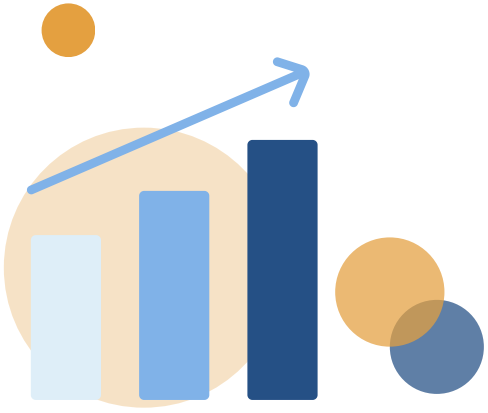
Los modelos de N-grama se utilizan para modelar la probabilidad de una palabra dado su contexto anterior en forma de N-1 palabras anteriores. Esto se conoce como la probabilidad condicional P(w | w_1, w_2, ..., w_(N-1)).

Los modelos de N-grama estiman estas probabilidades utilizando recuentos de ocurrencias de N-gramas en un corpus de texto de entrenamiento.

Desafíos y Limitaciones

Veamos algunos desafíos en los modelos N-grama:

Esparcidad de Datos



A medida que aumenta el tamaño del N-grama, la cantidad de datos necesarios para estimar con precisión las probabilidades también aumenta, lo que puede llevar a la esparcidad de datos en el corpus de entrenamiento.

Generalización Limitada

Los modelos de N-grama pueden tener dificultades para generalizar patrones más allá de los límites del tamaño de N, lo que puede afectar su capacidad para capturar la estructura semántica más amplia del lenguaje.

Sensibilidad al Tamaño del Corpus

La calidad y representatividad del corpus de entrenamiento pueden afectar significativamente el rendimiento de los modelos de N-grama.

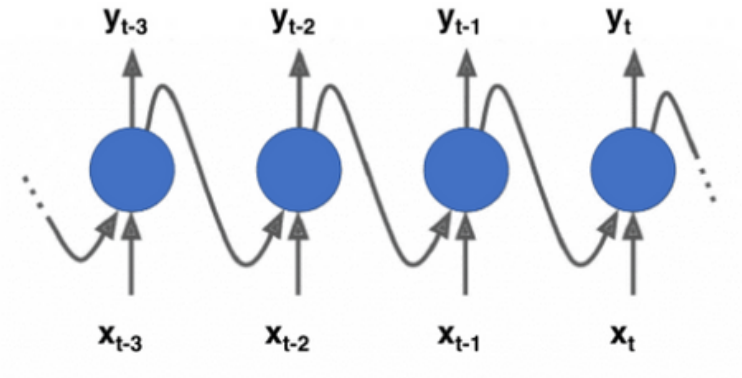
Los modelos de N-grama son una herramienta poderosa y versátil en el procesamiento del lenguaje natural, utilizados para modelar la estructura local del lenguaje y abordar una variedad de tareas, desde la generación de texto hasta la clasificación de documentos. Sin embargo, también tienen desafíos y limitaciones que deben considerarse al aplicarlos en diferentes contextos.

Modelos de Lenguaje Neuronales

Los modelos de lenguaje neuronales, como las redes neuronales recurrentes (RNN) y las redes neuronales transformadoras (Transformer), han ganado popularidad en los últimos años debido a su capacidad para capturar relaciones a largo plazo en secuencias de texto. Estos modelos aprenden representaciones vectoriales de palabras (embeddings) y utilizan capas recurrentes o de atención para procesar secuencias de manera más efectiva.

Modelo de lenguaje neuronal recurrente (RNN)

Las RNN son un tipo de red neuronal diseñada específicamente para procesar secuencias de datos. En el contexto del modelado de lenguaje, las RNN pueden capturar relaciones a largo plazo en secuencias de texto al mantener un estado interno que se actualiza con cada nueva palabra. Sin embargo, las RNN pueden sufrir el problema de desvanecimiento o explosión del gradiente.



Modelo de lenguaje basado en redes neuronales convolucionales (CNN)

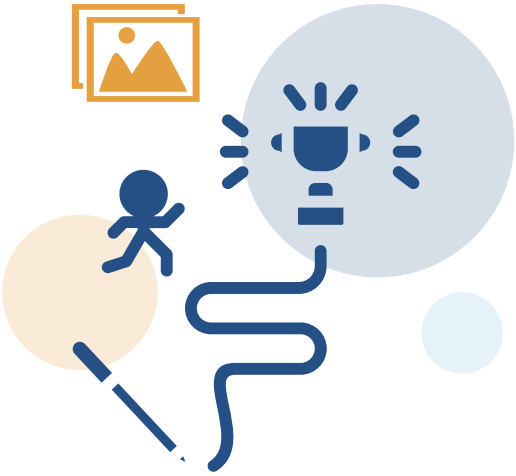
Aunque las CNN son más comunes en el procesamiento de imágenes, también se pueden aplicar al modelado de lenguaje. Estos modelos aplican operaciones de convolución sobre representaciones vectoriales de palabras para capturar características locales en el texto.

Modelo de lenguaje Transformer

Los modelos Transformer, introducidos por Vaswani et al. en 2017, han demostrado ser altamente efectivos en una variedad de tareas de NLP. Estos modelos utilizan una arquitectura completamente basada en atención y no requieren conexiones recurrentes. Son capaces de manejar relaciones a largo plazo en el texto y han sido adoptados en aplicaciones como la traducción automática y la generación de texto.

Modelo de lenguaje BERT (Bidirectional Encoder Representations from Transformers)

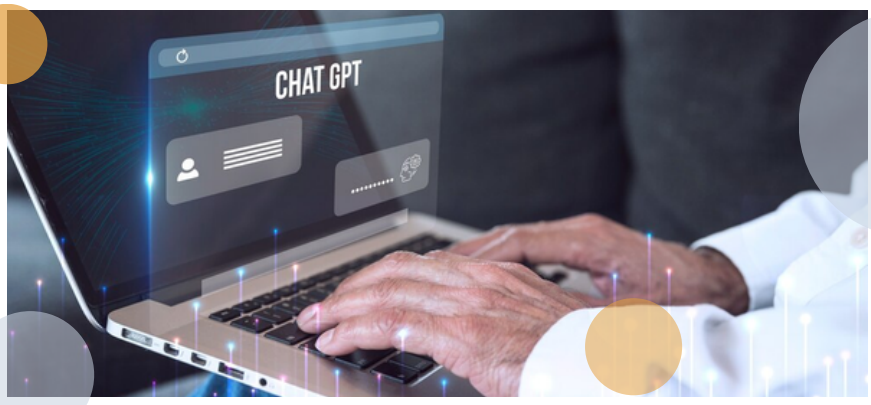
BERT es un modelo de lenguaje basado en Transformer pre-entrenado desarrollado por Google. Utiliza el aprendizaje de representaciones de lenguaje bidireccional para capturar el contexto de las palabras en una oración. BERT ha demostrado un rendimiento sobresaliente en una variedad de tareas de NLP, como la clasificación de texto, la respuesta a preguntas y la generación de texto.



2. Ejemplos de modelos de lenguaje comerciales y de código abierto

GPT (Generative Pre-trained Transformer) - OpenAI

Los transformadores generativos pre-entrenados (GPT) son un tipo de modelo de lenguaje grande (LLM) y un marco prominente para la inteligencia artificial generativa. Son redes neuronales artificiales que se utilizan en tareas de procesamiento del lenguaje natural. Los GPT están basados en la arquitectura de transformadores, pre-entrenados en grandes conjuntos de datos de texto no etiquetado, y capaces de generar contenido humano. A partir de 2023, la mayoría de los LLM tienen estas características y a veces se denominan ampliamente como GPTs.

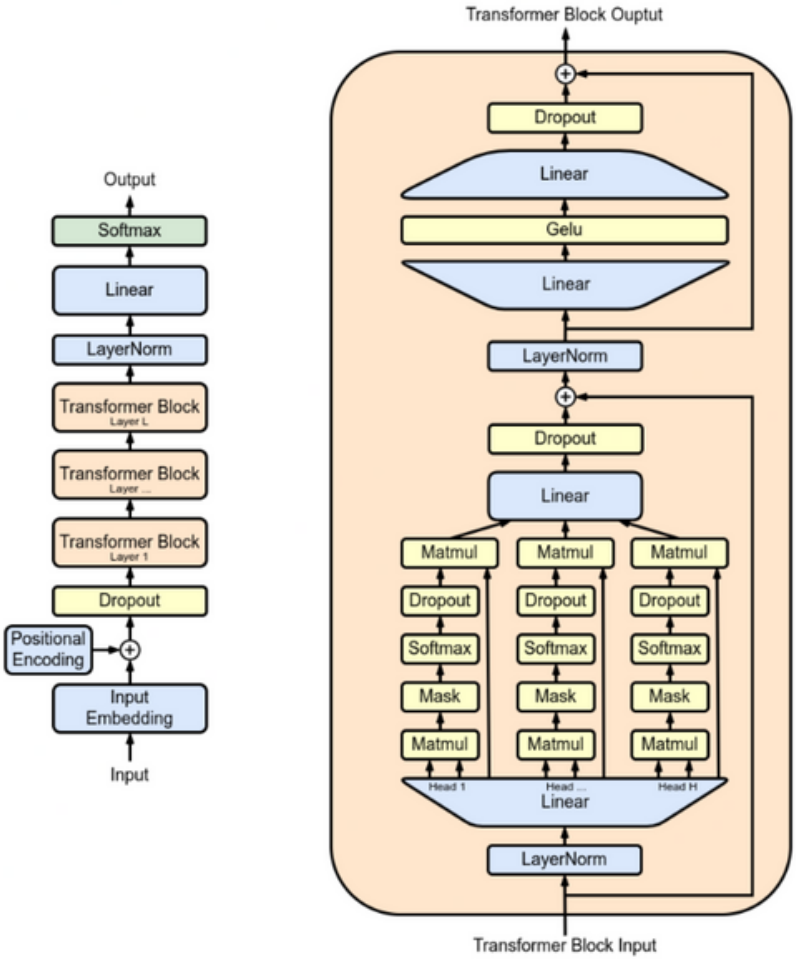


El primer GPT fue introducido en 2018 por OpenAI. OpenAI ha lanzado modelos fundamentales GPT muy influyentes que han sido numerados secuencialmente, para conformar su serie "GPT-n". Cada uno de estos fue significativamente más capaz que el anterior, debido al aumento de tamaño (número de parámetros entrenables) y entrenamiento. El más reciente de estos, GPT-4, fue lanzado en marzo de 2023.

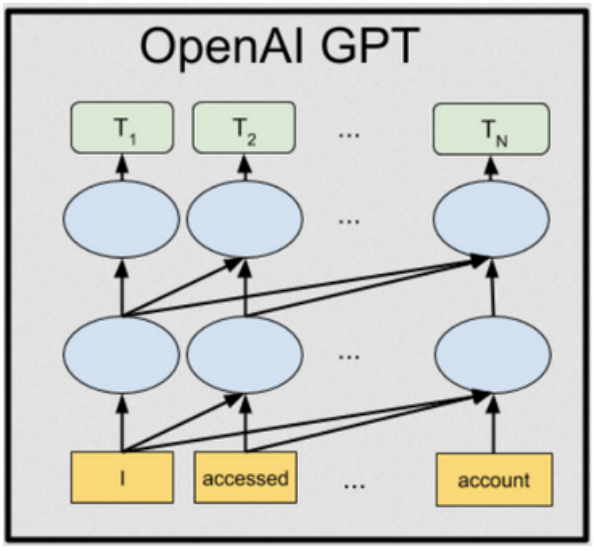
Tales modelos han sido la base para sus sistemas GPT más específicos para tareas, incluidos modelos ajustados para el seguimiento de instrucciones, que a su vez alimentan el servicio de chatbot ChatGPT.

El término "GPT" también se utiliza en los nombres y descripciones de tales modelos desarrollados por otros. Por ejemplo, otros modelos fundamentales GPT incluyen una serie de modelos creados por EleutherAI, y siete modelos creados por Cerebras en 2023. Además, empresas en diferentes industrias han desarrollado GPTs específicos para tareas en sus respectivos campos, como "EinsteinGPT" de Salesforce (para CRM) y "BloombergGPT" de Bloomberg (para finanzas).

1. Modelo GPT original



Estructura simplificada del modelo GPT



2. BERT (Bidirectional Encoder Representations from Transformers) - Google

Es un modelo de lenguaje basado en la arquitectura de transformadores, notable por su dramática mejora sobre los modelos anteriores de vanguardia. Fue introducido en octubre de 2018 por investigadores de Google. Una encuesta de literatura de 2020 concluyó que "en poco más de un año, BERT se ha convertido en un punto de referencia ubicuo en experimentos de Procesamiento del Lenguaje Natural (NLP), contando con más de 150 publicaciones de investigación que analizan y mejoran el modelo".

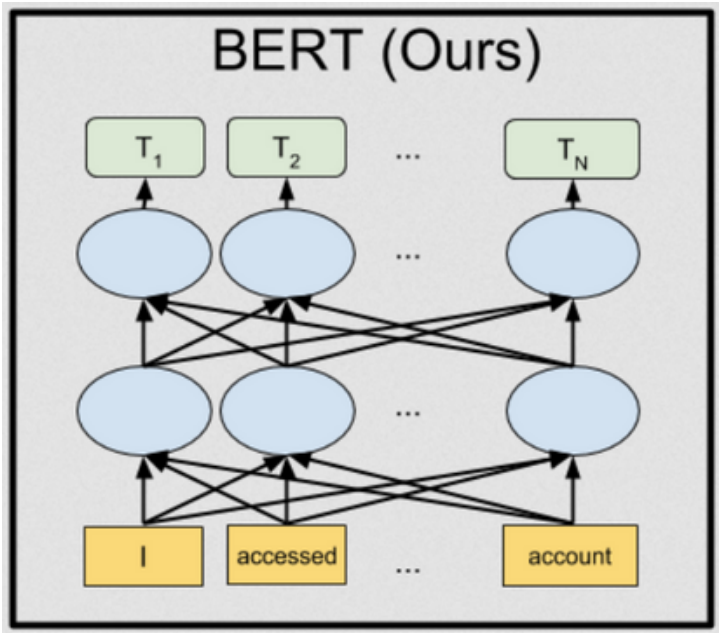
BERT fue implementado originalmente en el idioma inglés en dos tamaños de modelo:

BERTBASE: 12 codificadores con 12 cabezales de autoatención bidireccional, totalizando 110 millones de parámetros

BERTLARGE: 24 codificadores con 16 cabezales de autoatención bidireccional, totalizando 340 millones de parámetros.

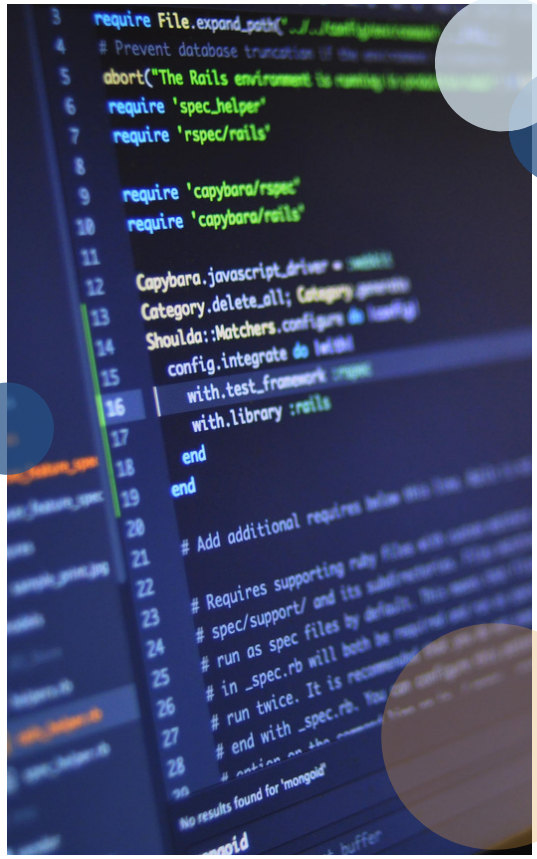
Ambos modelos fueron pre-entrenados en el Toronto BookCorpus (800 millones de palabras) y la Wikipedia en inglés (2,500 millones de palabras).

Estructura simplificada del modelo BERT



3. ELMo (Embeddings from Language Models)

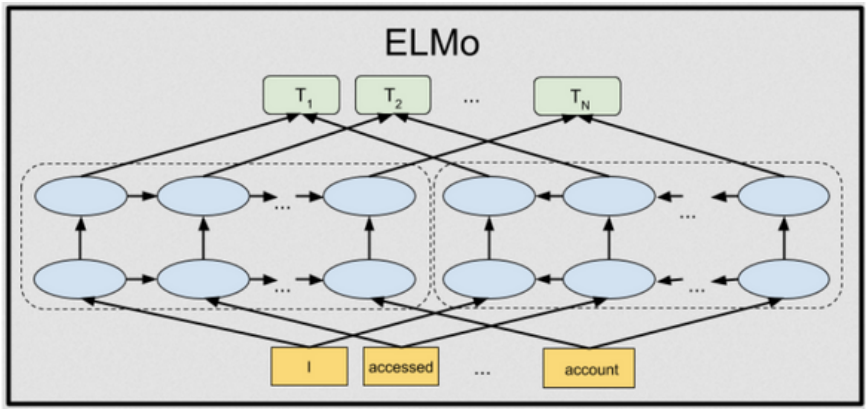
Es un método de incrustación de palabras para representar una secuencia de palabras como una secuencia correspondiente de vectores. Los tokens a nivel de caracteres se toman como entradas para un LSTM bidireccional que produce incrustaciones a nivel de palabras. Al igual que BERT (pero a diferencia de las incrustaciones de palabras producidas por los enfoques de "Bag of Words" y enfoques vectoriales anteriores como Word2Vec y GloVe), las incrustaciones de ELMo son sensibles al contexto, produciendo diferentes representaciones para palabras que comparten la misma ortografía pero tienen diferentes significados (homónimos) como "banco" en "banco de río" y "saldo bancario".



La innovación de ELMo proviene de su utilización de modelos de lenguaje bidireccionales. A diferencia de sus predecesores, estos modelos procesan el lenguaje en direcciones hacia adelante y hacia atrás. Al considerar todo el contexto de una palabra, los modelos bidireccionales capturan una comprensión más completa de su significado. Este enfoque holístico para la representación del lenguaje permite que ELMo codifique significados matizados que podrían pasarse por alto en modelos unidireccionales.

Fue creado por investigadores del Instituto Allen de Inteligencia Artificial y la Universidad de Washington y fue lanzado por primera vez en febrero de 2018.

Estructura simplificada del del modelo ELMo

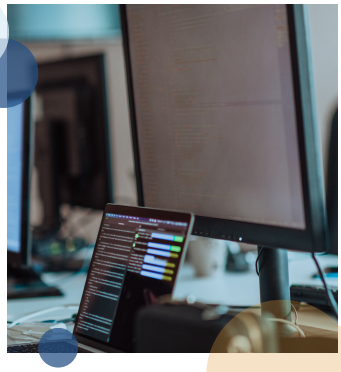


4. XLNet - Google/CMU

XLNet es un modelo de lenguaje pre-entrenado desarrollado por Google y la Universidad Carnegie Mellon. Utiliza una arquitectura Transformer extendida y utiliza una técnica llamada "permutación de segmento" para capturar relaciones entre todas las palabras en una oración. Ha demostrado un rendimiento competitivo en una variedad de tareas de NLP.

5. ULMFiT (Universal Language Model Fine-tuning) - Fast.ai

Desarrollado por Fast.ai, ULMFiT es un enfoque para el entrenamiento de modelos de lenguaje que permite el ajuste fino en tareas específicas con conjuntos de datos pequeños. Utiliza una arquitectura LSTM y ha sido ampliamente utilizado en la comunidad de aprendizaje automático para una variedad de aplicaciones de procesamiento de texto.



Estos son solo algunos ejemplos de modelos de lenguaje disponibles comercialmente o de código abierto. Cada modelo tiene sus propias características y aplicaciones específicas, y la elección del modelo depende de los requisitos y objetivos del proyecto de NLP.