Modelos Fundacionales: Concepto y Paradigma de los Nuevos Modelos de Gran Tamaño

Miguel Cotrina

Miguel Cotrina

- **Perfil Académico**
  - Ingeniero de software por la Universidad Tecnológica del Perú.
  - Maestría en ciencia de datos por la Universidad Ricardo Palma.
- **Perfil Profesional**
  - Arquitecto de datos en Clínica Internacional
  - Consultor e Instructor de Big Data, cloud, IA e IA Gen en empresas privadas y públicas

Agenda

- Introducción teórica
- IA Generativa Multimodal
- Modelos on-site y cloud

Como comenzo “Attention is All You Need”

Introducida en el artículo "Attention is All You Need" realizado por investigadores de Google en el año 2017, la arquitectura Transformer está diseñada para procesar y generar texto muy similar al humano para una amplia gama de tareas o procesos, desde la traducción automática hasta la generación de texto de propósito general.

https://arxiv.org/abs/1706.03762

Foundation models

Los foundation models son sistemas de IA entrenados con grandes volúmenes de datos no etiquetados, adaptables a múltiples tareas. El término fue acuñado en 2021 por el Centro de Investigación de Modelos Fundacionales (CRFM) de Stanford, elegido en lugar de "modelo de lenguaje grande" porque su enfoque va más allá del lenguaje, abarcando diversas aplicaciones generales.

https://crfm.stanford.edu/workshop.html

Foundation models

Estados Unidos: La Orden Ejecutiva sobre el Desarrollo y Uso Seguro, Confiable y Responsable de la Inteligencia Artificial define un modelo fundacional como "un modelo de IA entrenado con datos amplios; que generalmente utiliza autosupervisión; contiene al menos decenas de miles de millones de parámetros; y es aplicable en una amplia gama de contextos".

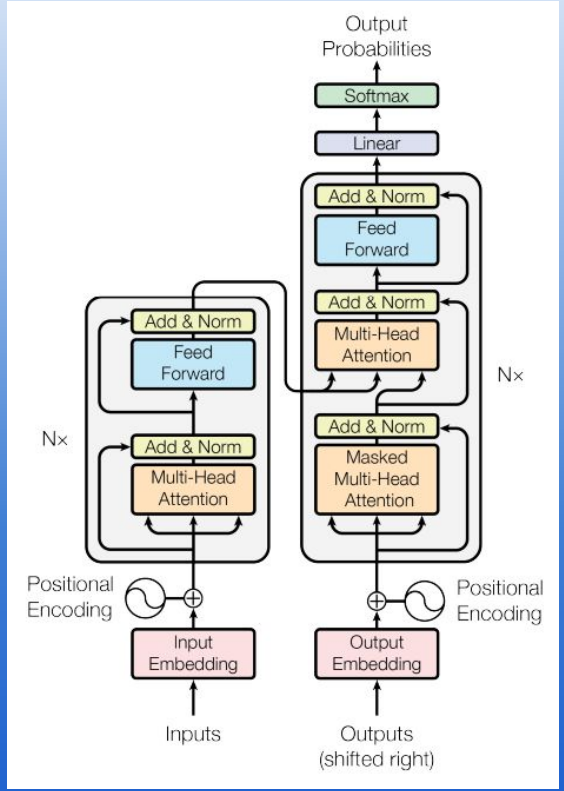
https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

Foundation models

Unión Europea: La posición negociada del Parlamento Europeo sobre la Ley de IA define un modelo fundacional como un "modelo de IA entrenado con datos amplios a escala, diseñado para la generalidad de resultados, y que puede adaptarse a una amplia gama de tareas distintivas".

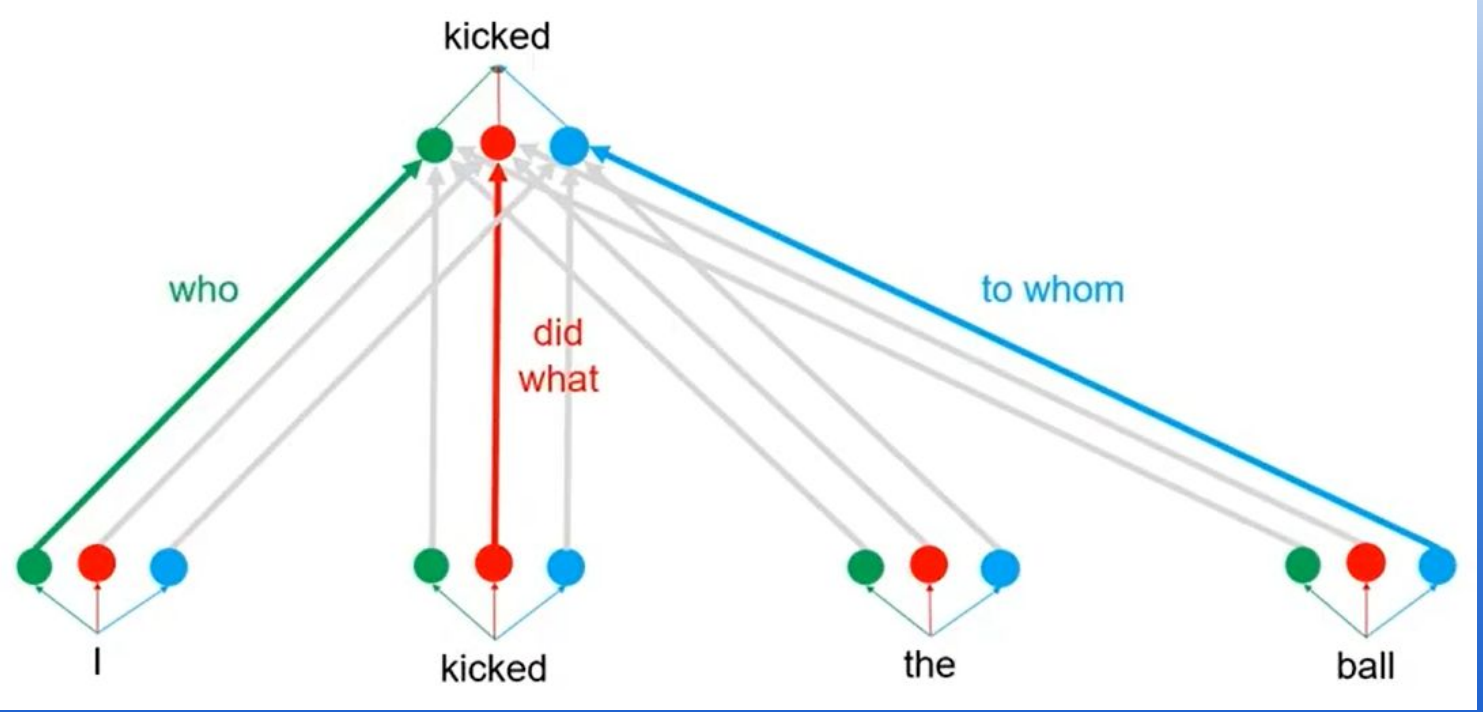
Arquitectura transformers

Los Transformers revolucionaron el análisis del lenguaje procesando todas las palabras de una frase simultáneamente, mejorando velocidad y comprensión del contexto, incluso con palabras distantes. Su avance clave, el mecanismo de autoatención, identifica la relevancia de cada palabra en relación con las demás



https://www.truefoundry.com/blog/transformer-architecture

Arquitectura transformers



Miguel Cotrina

https://www.youtube.com/watch?v=2yb186AfjX0

Conclusion

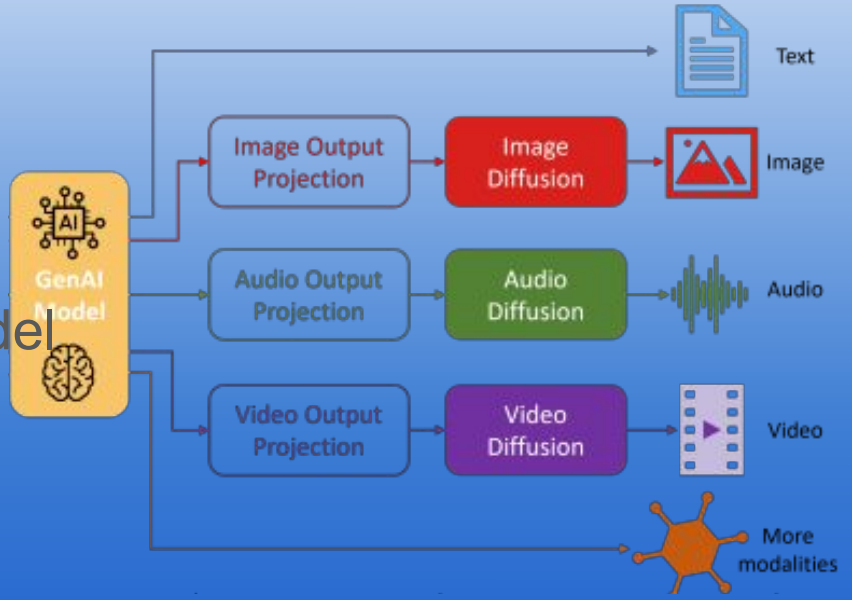
En resumen, los foundation models y los Transformers representan hitos clave en la evolución de la inteligencia artificial. Destacan por su versatilidad al adaptarse a múltiples tareas gracias a su entrenamiento con grandes volúmenes de datos no etiquetados, superando términos previos como "modelos de lenguaje grande". Por otro lado, los Transformers, con su mecanismo de autoatención, han revolucionado el procesamiento del lenguaje, mejorando la velocidad y comprensión contextual, marcando un estándar en aplicaciones como traducción y generación de texto.

Inteligencia artificial generativa

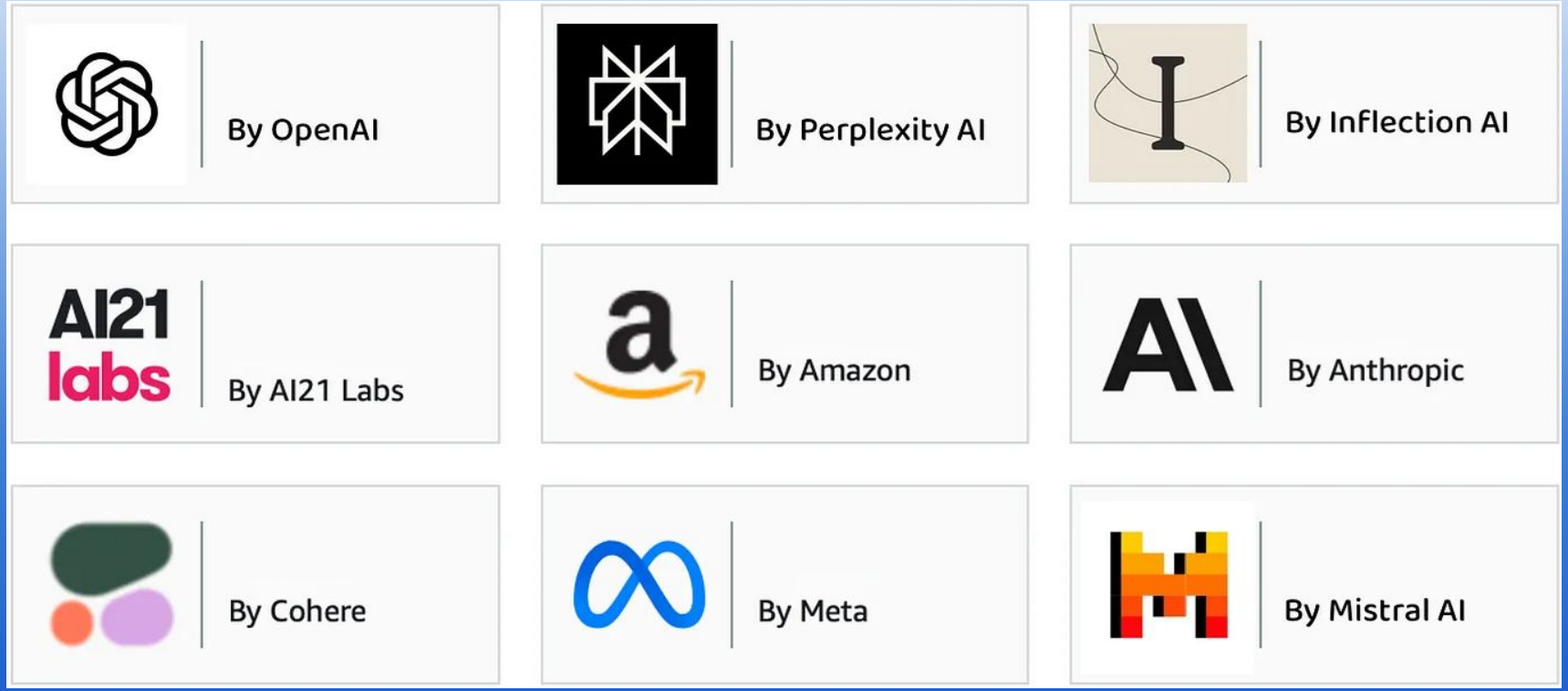
Se le llama inteligencia artificial generativa porque estas tecnologías tienen la capacidad de crear contenido nuevo a partir de patrones y datos preexistentes. A diferencia de otros tipos de IA que solo analizan o clasifican información

Inteligencia artificial generativa multimodal

Estos modelos pueden generar contenido en múltiples formatos, como texto, imágenes, audio o video, a partir de datos de diferentes tipos o combinaciones de ellos. La palabra "multimodal" se refiere a la capacidad del modelo de procesar y relacionar información de diversas fuentes, como texto e imágenes, para producir resultados integrados y coherentes.



Principales empresas



Principales Modelos on-site y Cloud Destacados

Modelos disponibles en línea:

- **ChatGPT (GPT-4)**: Desarrollado por OpenAI, accesible a través de su plataforma web y API, destacando en interacción conversacional y generación de contenido.
- **Google Bard (PaLM 2)**: Modelo de Google enfocado en generación de texto coherente y relevante, con capacidades multimodales en desarrollo.
- **Claude 3**: Creado por Anthropic, este modelo prioriza la seguridad y alineación con valores humanos, ideal para aplicaciones conversacionales.
- **Gemini (Google DeepMind)**: Modelo multimodal avanzado anunciado en 2023, competidor directo de GPT-4, diseñado para tareas complejas de generación y análisis.

Principales Modelos on-site y Cloud Destacados

Modelos instalables en máquinas con GPU:

- **LLaMA 3:** Modelo de código abierto de Meta, flexible y eficiente para implementaciones locales y personalización.
- **GPT-NeoX-20B:** Modelo de código abierto desarrollado por EleutherAI, adecuado para aplicaciones locales con soporte de GPU.
- **Gemma (Google):** Familia de modelos ligeros de código abierto, diseñados para ofrecer alto rendimiento en dispositivos móviles, hardware propio y servicios alojados.

Laboratorio

Codigo compartido

- Implementación en Colab de Gemma y Llama
- Implementación de GPT-4 en Colab

Enlace de materiales:

https://github.com/macespinoza/programa7genai/tree/main/Clase%2001

Agradecimiento y preguntas

Muchas gracias a todas las personas que están interesadas en aprender sobre estas nuevas tecnologías, el camino comienza pero el destino aún es desconocido.

Todas sus preguntas consultas o feedback son bienvenidos y lo pueden dejar en los comentarios del video de cada clase

Redes sociales:

- https://www.linkedin.com/in/mcotrina/
- https://www.youtube.com/@macespinozaonline
- https://github.com/macespinoza/

Programa de Introducción a la IA Generativa con Modelos de Gran Tamaño de 7 clases

Miguel Cotrina