

Inteligencia Artificial



KEEPCODING
Tech School



Contenido de la Masterclass

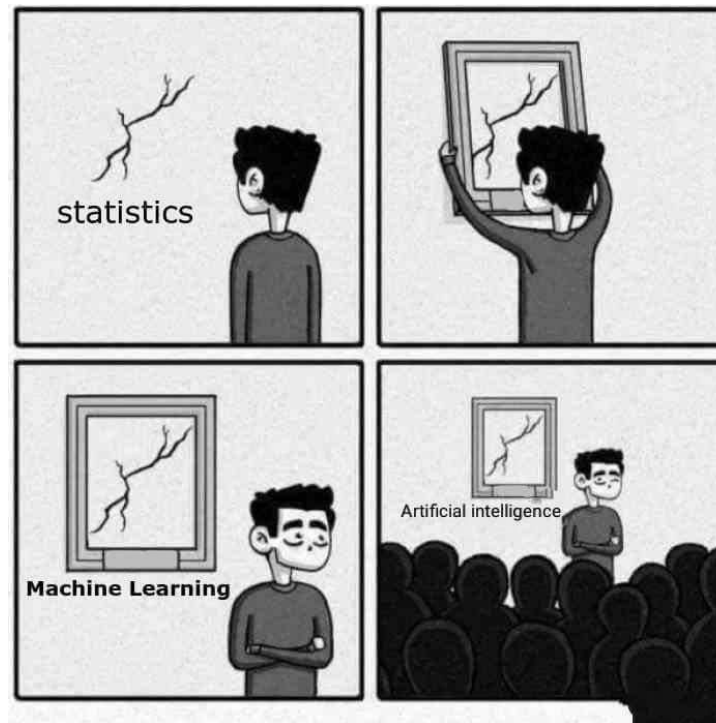
- Que es un LLM
- Transformers
- Prompt engineering
- Generative configuration
- Pre-training LLMs
- Fine-tuning
- FLAN
- Model evaluation
- Como sigue?



INTRODUCCIÓN

Machine Learning

- Algoritmo que en vez de seguir instrucciones concretas, aprende de los datos
- Hace predicciones o toma decisiones a partir de su aprendizaje
- Un algoritmo de ML es tan bueno como buenos son su datos



Modelos discriminativos

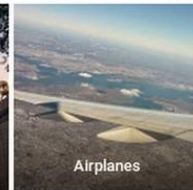
- Aprende a diferenciar entre categorías o grupos basándose en los datos de entrenamiento
- Clasificación de imágenes, detección de objetos, modelado de lenguaje y más.
- SVM, Random Forest, Redes Neuronales profundas,...

Google Mistakenly Tags Black People as 'Gorillas,' Showing Limits of Algorithms



diri noir avec banan @jackyalcine · Jun 28

Google Photos, y'all [redacted] up. My friend's not a gorilla.



RETWEETS

1,031

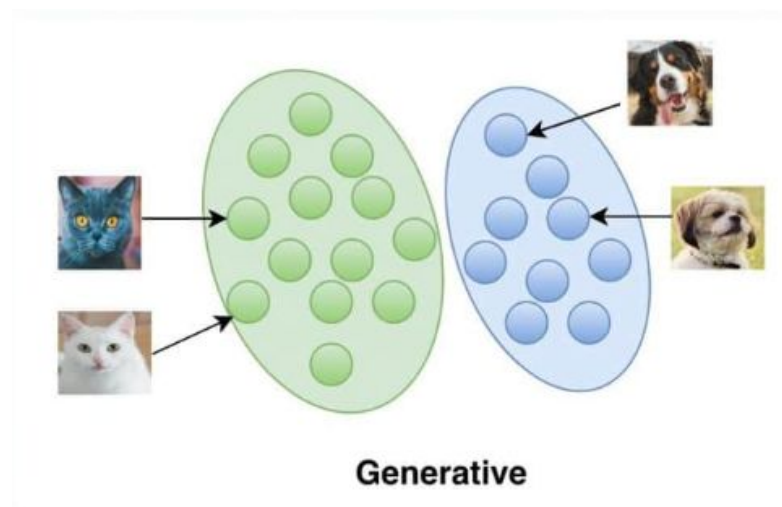
FAVORITES

513



Modelos generativos

- Generar nuevas muestras a partir de datos de entrenamiento existentes
- También puede clasificar
- Naive Bayes, Hidden Markov Models, Autoencoder, Boltzmann Machines, Variational Autoencoder, Generative Adversarial Networks



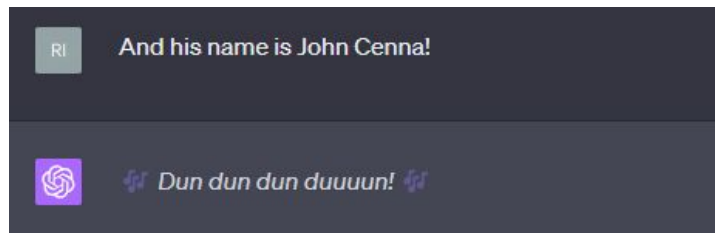
IA Generativa

- Modelos de IA que crean contenido nuevo (imágenes, texto, música)
- Generan el contenido a partir de datos de entrenamiento
- Arte, diseño, simulación, programación, entretenimiento...
- Entrenamiento complejo, variabilidad en resultados y posibles usos éticamente cuestionables.



Large Language Models

- Una rama de la IA que imita la generación humana de contenido. Se entrenan con billones de palabras durante meses
- Con miles de millones de parámetros, estos modelos descomponen tareas complejas, razonan y resuelven problemas.



LLMs

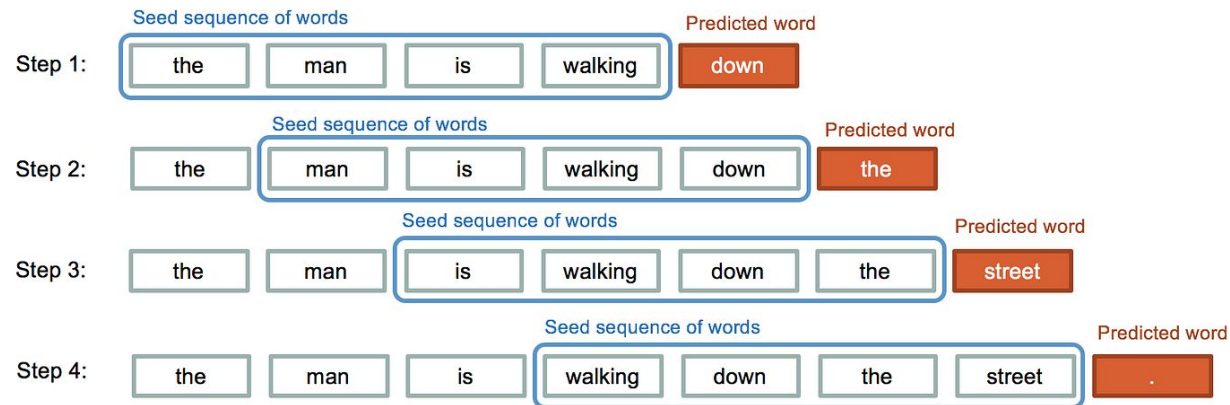
El Poder de los LLMs

- Chatbots avanzados con lenguaje natural
- Versatilidad:
 - Escritura de textos a partir de indicaciones.
 - Resumen de diálogos y conversaciones.
 - Traducción entre idiomas y a código de programación.
 - Recuperación de información específica, como reconocimiento de entidades
- Conexión con Fuentes Externas:
 - Integración con APIs y bases de datos.
 - Ampliación del conocimiento del modelo más allá de su entrenamiento.

Generación de texto antes de los transformers

- Redes Neuronales Recurrentes (RNN)
- Tipo especial de red neuronal diseñada para reconocer patrones en secuencias de datos.
- Poseen "memoria" para recordar entradas anteriores.
- Problema del desvanecimiento del gradiente: dificulta el aprendizaje de dependencias a largo plazo.
- Computacionalmente intensivas con secuencias largas.

Generación de texto antes de los transformers





TRANSFORMERS

La revolución de los Transformers

- Avance sobre las RNNs en tareas de procesamiento de lenguaje natural.
- Capacidad para aprender la relevancia y el contexto de todas las palabras en una frase.
- Uso de pesos de atención para aprender relaciones intrínsecas entre palabras.



Estructura general de un transformer

- Revolución
- Google
- 2017

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

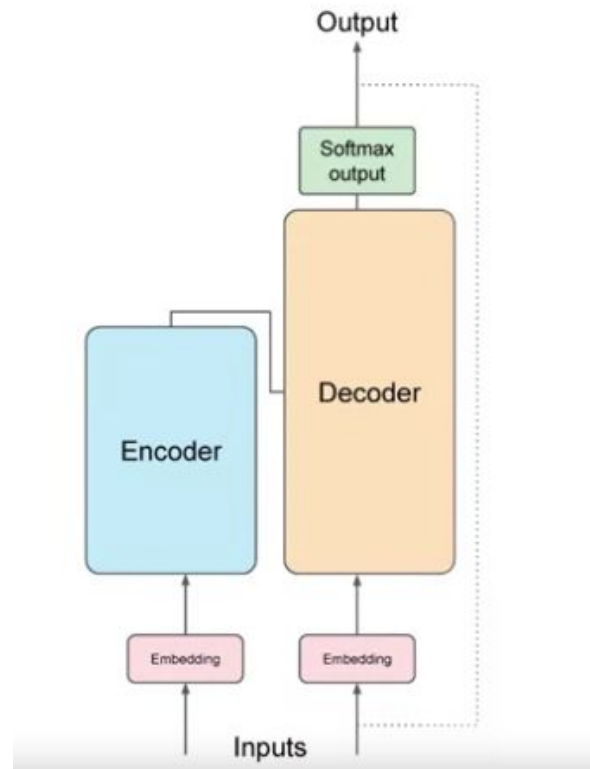
Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Lukasz Kaiser*
Google Brain
lukaszkaiser@google.com

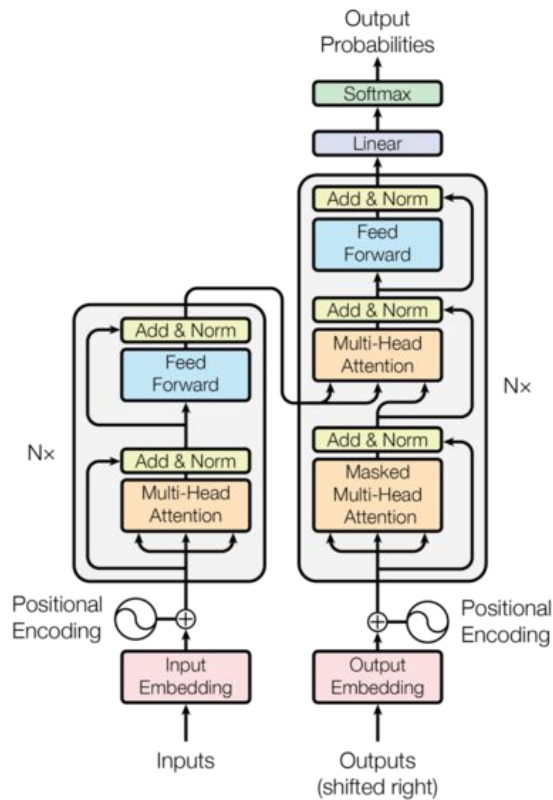
Illia Polosukhin* ‡
illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.



Estructura general de un transformer



Relevancia y Contexto de las Palabras

- Los Transformers capturan la relación de cada palabra con cada otra palabra en una oración.


Soy uno con la fuerza y la fuerza está conmigo.

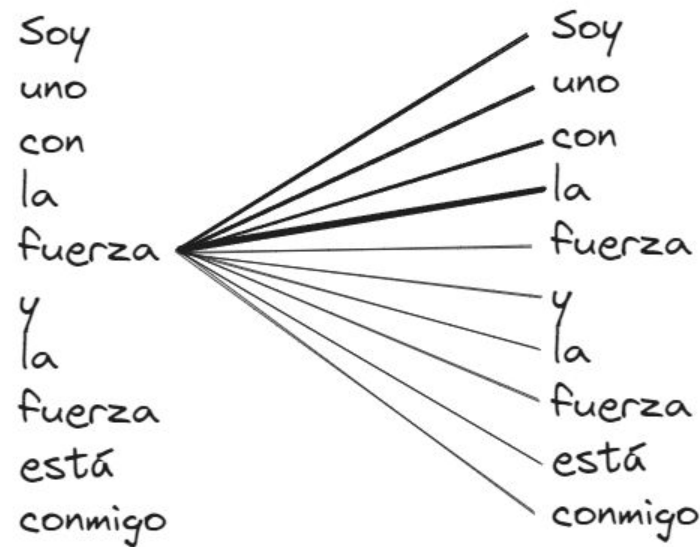
RNN


Soy uno con la fuerza y la fuerza está conmigo.


Transformers

El poder de la atención

- Aplicación de pesos para aprender relevancias
- Capacidad de aprender en la atención de toda la frase



Tokenización

GPT-3 Codex

Soy uno con la fuerza y la fuerza esta conmigo

Clear

Show example

Tokens

19

Characters

46

Soy uno con la fuerza y la fuerza esta conmigo

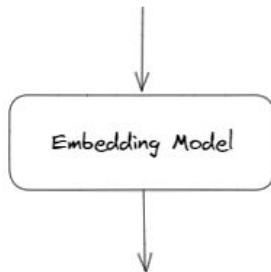
[50, 726, 555, 78, 369, 8591, 14035, 263, 4496, 331, 8591, 14035, 263, 4496, 1556, 64, 369, 76, 14031]

<https://platform.openai.com/tokenizer>

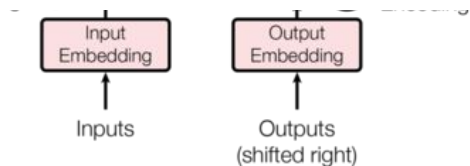
Incrustación (Embedding)

- Convierte palabras en vectores numéricos
- Capturan el contexto y la semántica de las palabras
- Se ha entrenado previamente
- La proximidad o distancia entre vectores en este espacio puede interpretarse como similitud o diferencia semántica.

Soy uno con la fuerza y la fuerza esta conmigo
[50, 726, 555, 78, 369, 8591, 14035, ...]



[[0.12, 0.85], [-0.58, 0.39], [0.32, -0.21], ...]



Positional encoding

- Codificación posicional para mantener el orden de las palabras
- Estas incrustaciones posicionales se suman a las incrustaciones de palabras.

$[[0.12, 0.85], [-0.58, 0.39], [0.32, -0.21], \dots]$

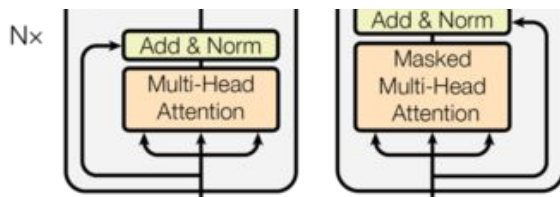
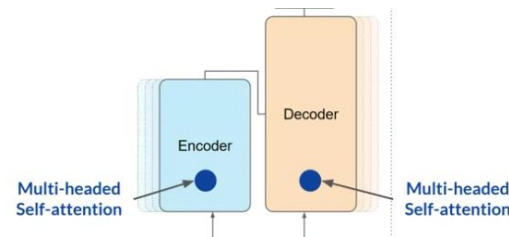


$\begin{bmatrix} [0][0.12, 0.85], \\ [1][-0.58, 0.39], \\ [2][0.32, -0.21], \\ \dots \end{bmatrix}$



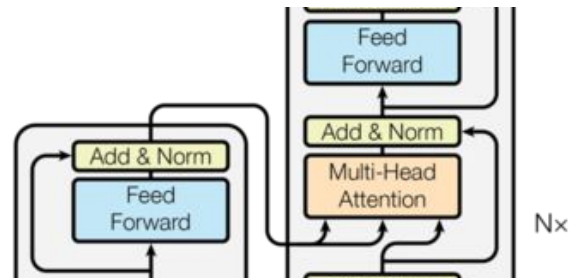
Multihead attention

- El modelo presta “atención” a diferentes partes de la información
- Una head podría centrarse en las relaciones entre las entidades, conectando "fuerza" con "Soy".
- Otra head podría centrarse en la acción, conectando "uno" con "con".
- Una tercera "head" podría centrarse en otros aspectos, como la estructura gramatical.



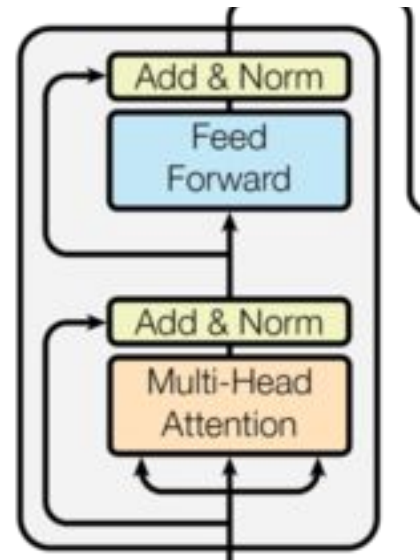
Red feedforward

- Una vez que se aplican los pesos de atención, la información se procesa y se traduce en predicciones para el siguiente token
- Ajusta y refina la comprensión por atención



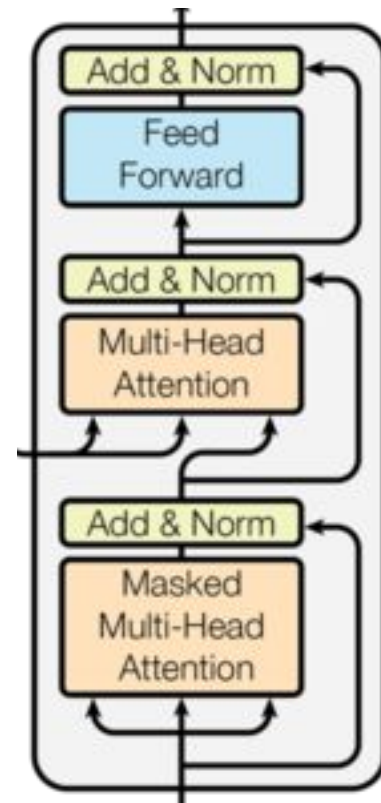
Encoder

- Transforma la entrada en una representación abstracta, codificando las relaciones entre sus partes.
- Esta representación es un conjunto de vectores que capturan la esencia de la entrada

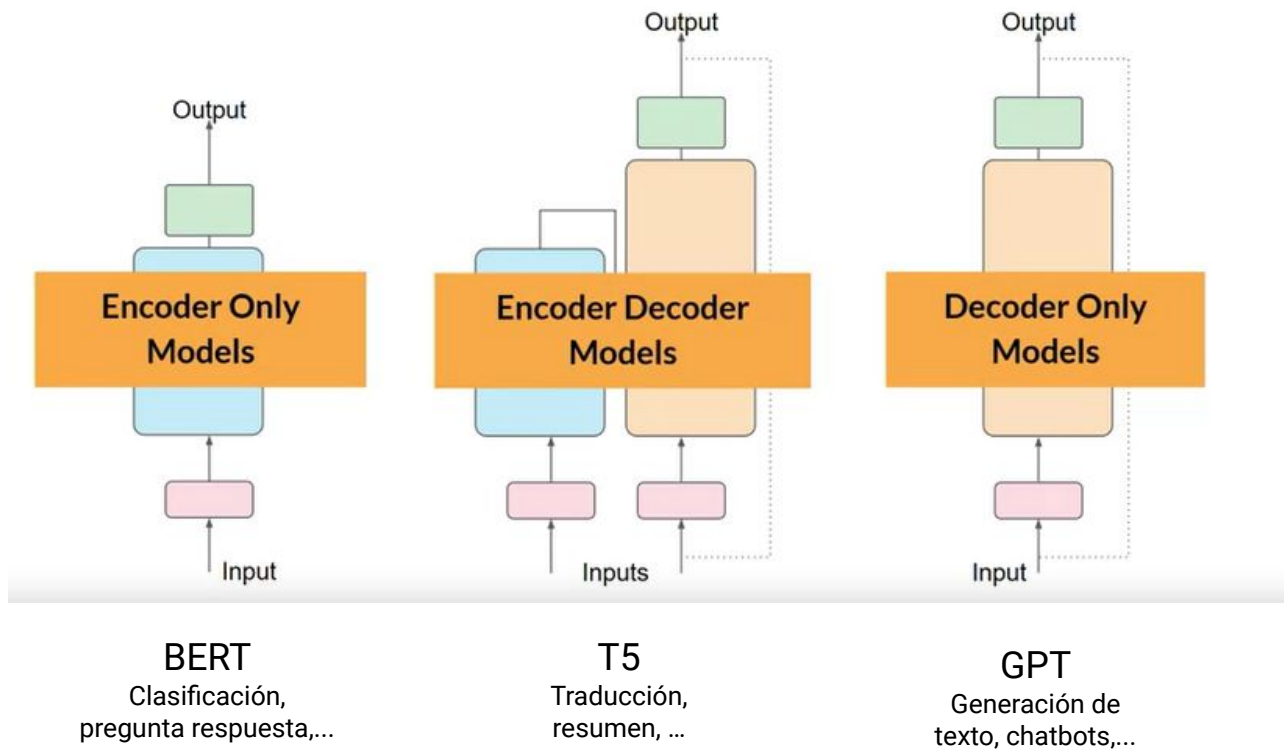


Decoder

- Toma la representación intermedia del codificador y la "interpreta" para generar una secuencia de salida
- En ChatGPT, el decodificador toma tu mensaje y genera una respuesta basada en su entrenamiento previo.



Transformers



En resumen

- La arquitectura "Transformer" es un modelo para datos secuenciales que utiliza "atención" para procesar palabras simultáneamente y capturar relaciones entre ellas.
- El "Encoder" es como un lector que comprende y resume una historia.
- El "Decoder" toma ese resumen y lo traduce o cuenta de una forma diferente o en otro idioma.



Más información



<http://jalammar.github.io/illustrated-transformer/>



Prompt engineering





Terminología Básica y Desafíos

- Prompt, Inferencia, Completion, Context Window.
- Los modelos no siempre aciertan al primer intento.
- El arte de mejorar el prompt: **Prompt Engineering**.



InContext Learning y Tipos de Inferencia

- Mejorar el entendimiento del modelo con ejemplos.
- Zero-shot, One-shot, Few-shot.
- Los ejemplos clarifican y demuestran la tarea al modelo.



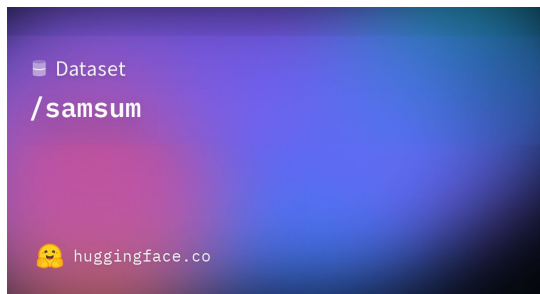
Generative Configuration



Generative Configuration

- **Max New Tokens:** Limita la cantidad de tokens que el modelo generará.
- **Temperatura:** Ajusta la distribución de probabilidad para el próximo token, afectando la creatividad del texto generado.

LAB 1





Pre-training LLMs

Desafíos al entrenar LLMs

- Los LLMs son enormes
- Necesidad de gran memoria para almacenar y entrenar sus parámetros.



RAM

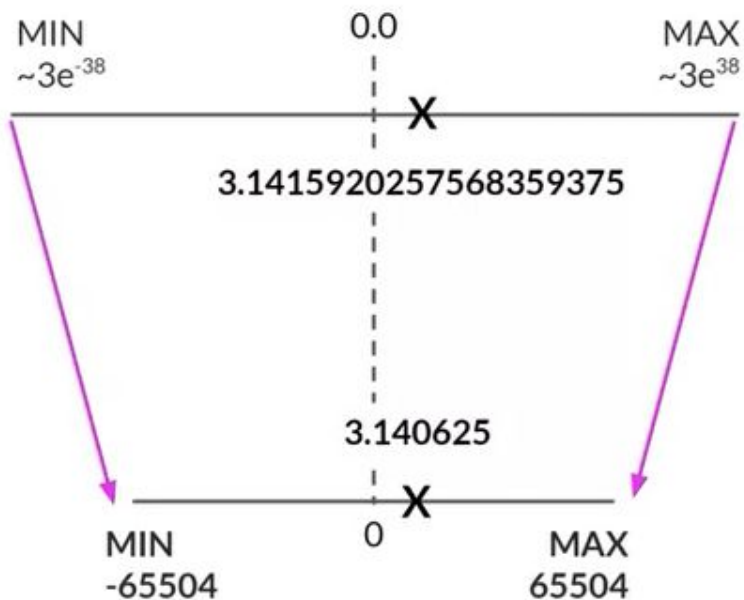
- Parámetro = Pesos configurados en cada neurona de la RN
- Memoria necesaria para guardar el modelo a 32 bits full precision
 - 1 parámetro = 4 bytes
 - 1B parámetros = $4 \times 10^9 = 4\text{GB}$
- Memoria necesaria para entrenar el modelo
 - 1 parámetro = 4B + 8B (optimizador) + 4B (gradiente) + 8B (activaciones) = 24B
 - 1B parámetros = 80GB



Cuantificación

- Reducir la memoria RAM necesaria
- Proyecta número de 32 bits a 16 bits o 8 bits
- 75% de ahorro y aumento en la velocidad
- Pérdida de precisión

Cuantificación



Let's store Pi: 3.141592

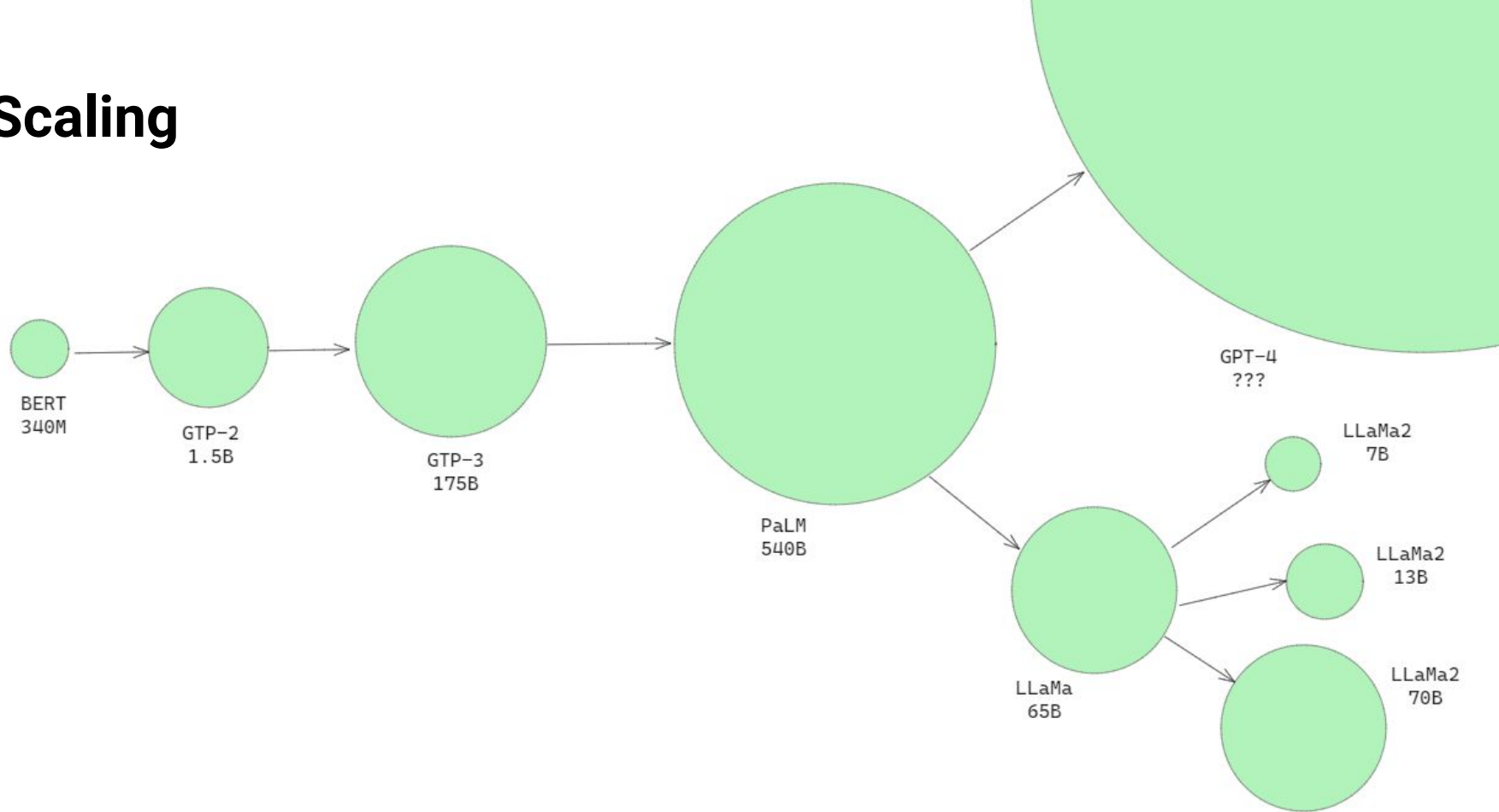
FP32

0	10000000	10010010000111111011000
<hr/>		
Sign 1 bit	Exponent 8 bits	Fraction 23 bits

FP16

0	10000	1001001000
<hr/>		
Sign 1 bit	Exponent 5 bits	Fraction 10 bits

Scaling





Fine-tuning

Fine-tuning

- Mejorar los LLM para que hagan tareas específicas mejor
- No hace falta añadir información (prompting)
- El modelo aprende nuevos datos

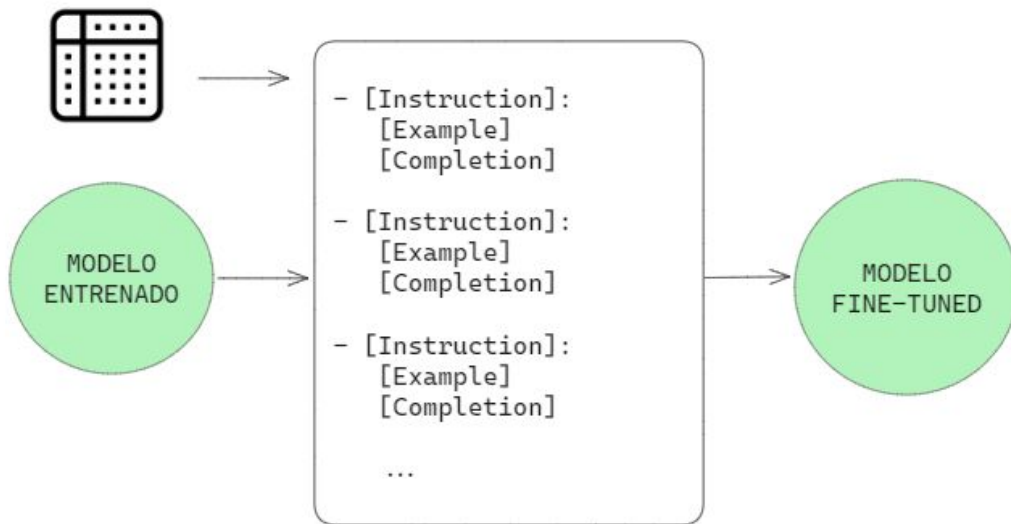


Instruction fine-tuning

- Entrenamiento de un modelo para responder a comandos lingüísticos específicos.
- Permite un modelo versátil para diversas tareas basadas en instrucciones del usuario.
- Ideal para modelos que necesitan realizar variadas funciones a partir de indicaciones claras.

Instruction fine-tuning

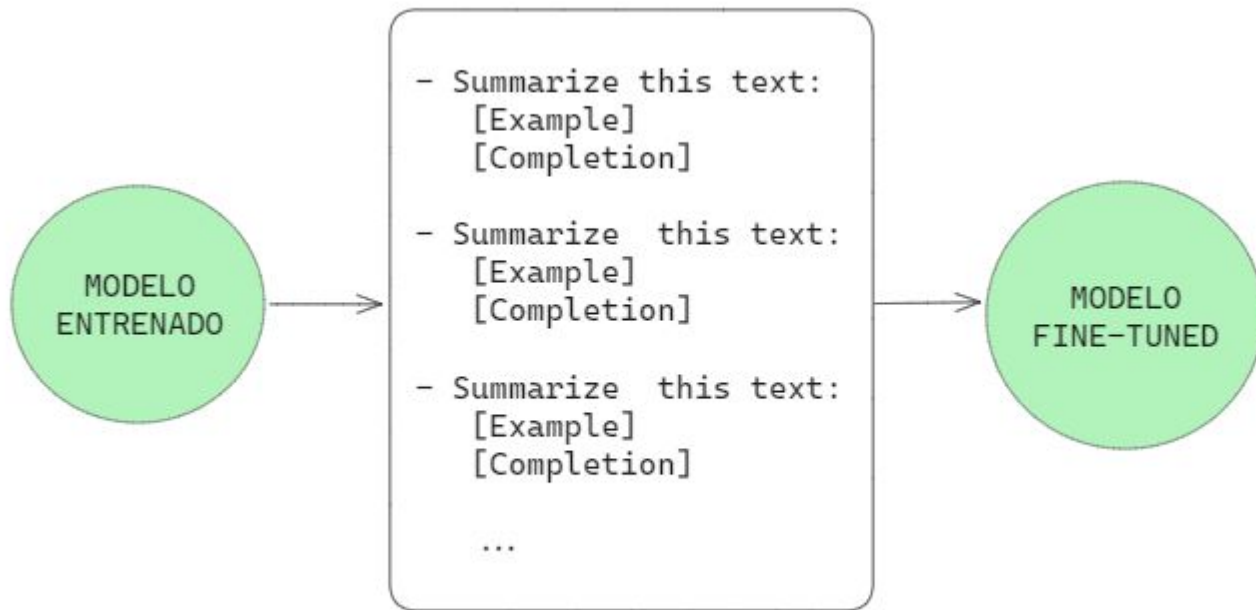
Instructions dataset



Single task fine-tuning

- Mayor precisión en tareas específicas.
- Se pueden obtener buenos resultados con pocos ejemplos (500-1000)
- Riesgo de "Olvido Catastrófico".
- Puede degradar el rendimiento en otras tareas.
- El "Olvido Catastrófico" ocurre cuando el modelo pierde la capacidad de realizar tareas previamente aprendidas al ser afinado para una nueva tarea.

Single task fine-tuning

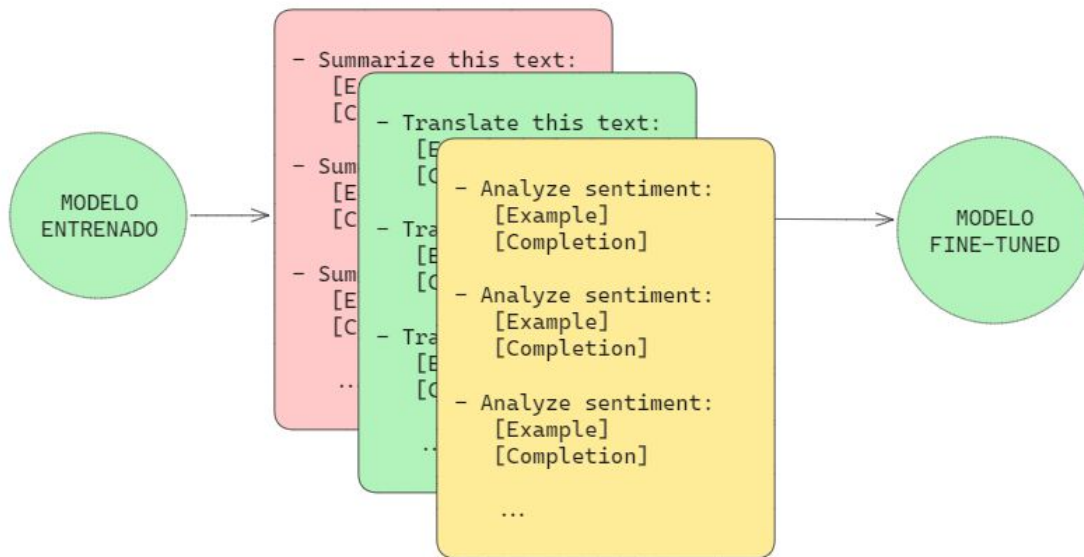




Multi-task fine-tuning

- Reduce el riesgo de "Olvido Catastrófico".
- Versatilidad en múltiples tareas.
- Requiere una gran cantidad de datos (50-100,000 ejemplos).
- Mayor demanda computacional.

Multi-task fine-tuning



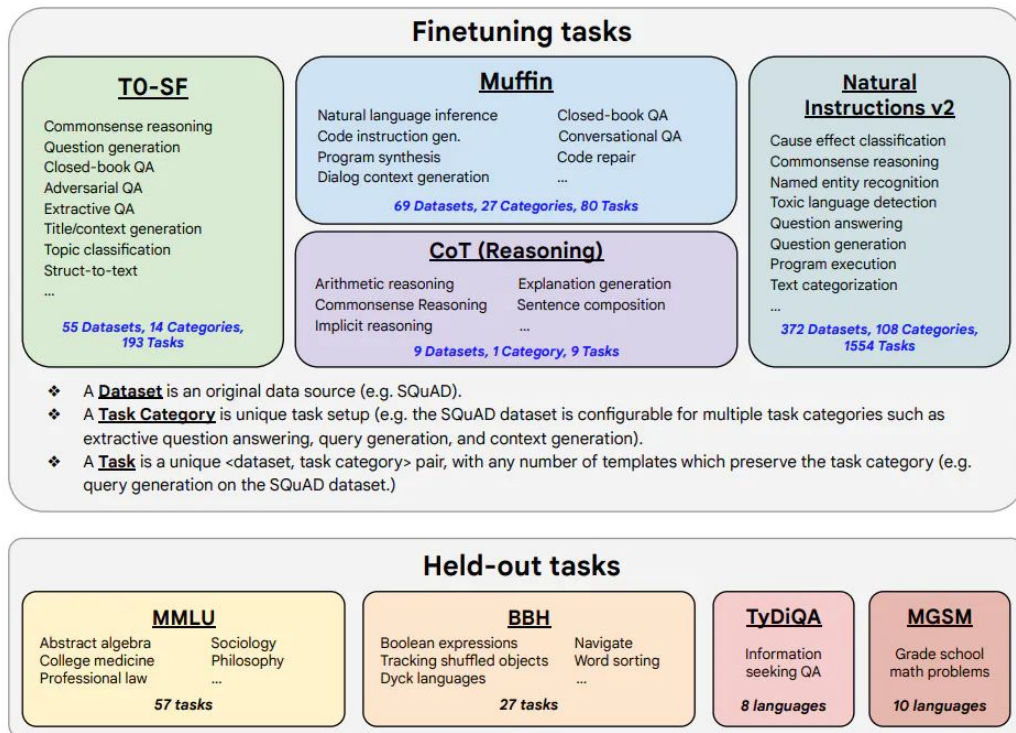


FLAN

Fine-tuned LAnguage Net

- Conjunto específico de instrucciones utilizadas para afinar diferentes modelos
- Ejemplos incluyen FLAN-T5 (versión instruida de T5) y FLAN-PALM (versión instruida del modelo PALM)
- FLAN-T5 ha sido afinado en 473 conjuntos de datos a través de 146 categorías de tareas.

Fine-tuned LAnguage Net



Scaling Instruction-Finetuned Language Models



Model evaluation

Model evaluation

- Mi modelo hace su trabajo correctamente?

$$\text{Precisión} = \frac{\text{Predicciones correctas}}{\text{Predicciones totales}}$$

Luke, yo soy tu padre

Luke, yo mate a tu padre





Métrica ROUGE

- Recall **O**riented **U**nder study for **J**esting **E**valuation
- Evalúa la calidad de resúmenes generados automáticamente.
- Compara el resumen de la máquina con un resumen de referencia humano.

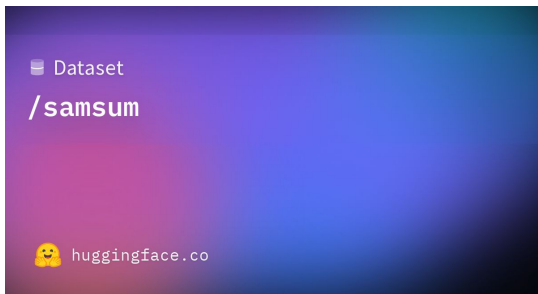


Métrica BLEU



- **Bi**Lingual **E**valuation **U**nderstudy
- Evalúa la calidad de textos traducidos por máquinas.
- Confronta una traducción automática con una traducción de referencia creada por humanos.

LAB 2





¡Esto sigue!





Falcon 180B



KEEPCODING

Tech School

Madrid | Barcelona | Bogotá

Eric Risco de la Torre

erisco@icloud.com

<https://www.linkedin.com/in/erisco-and/>