Procesamiento del Lenguaje Natural (NLP). Large Language Models: Familias de Transformers

Inteligencia Artificial e Ingeniería del Conocimiento

Constantino Antonio García Martínez

Universidad San Pablo CEU

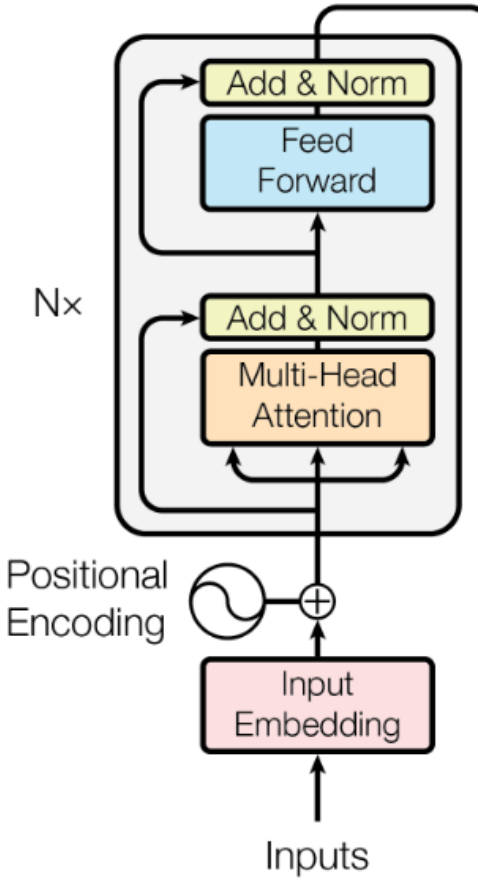
Bibliografía

- Vaswani et al. Attention is all you need. Advances in neural information processing systems. 2017.

Attention is All you Need: Transformers

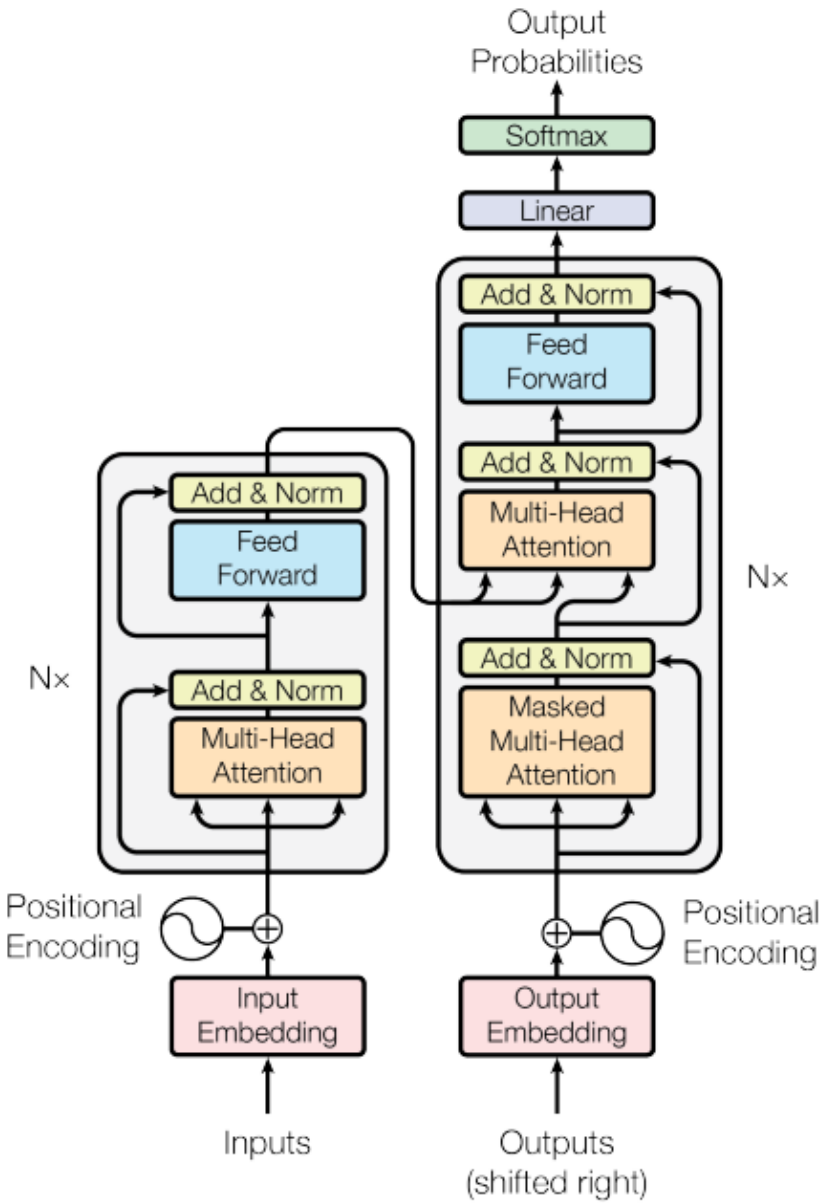
Attention is All you Need: Transformer-Encoder

- En la anterior sesión estudiamos **parte** de la arquitectura **Transformer** introducida en “Attention is all you need"(Vaswani et al., 2017).
- Nos centramos en el **Encoder** de la arquitectura original:



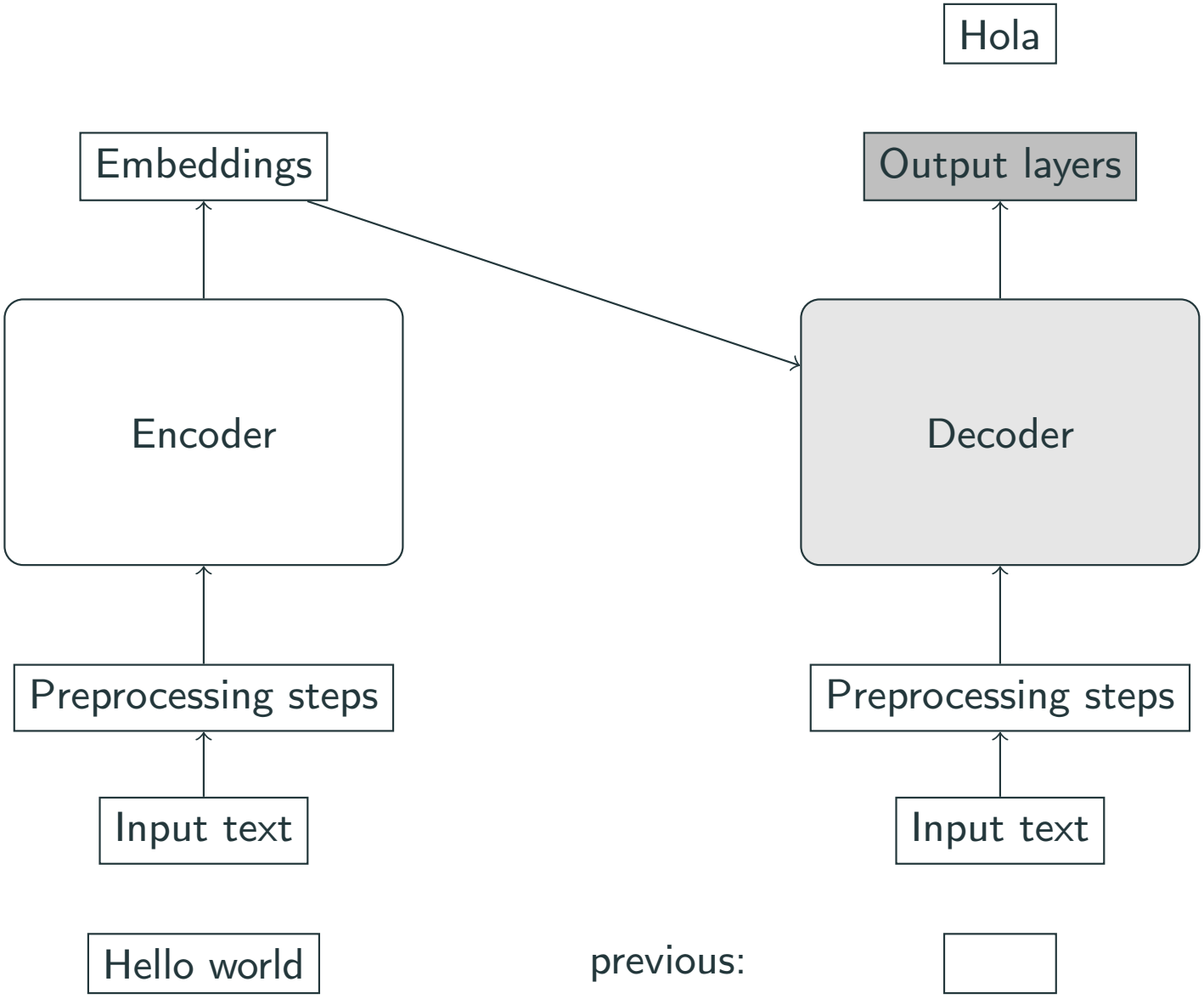
Attention is All you Need: Transformer Completo

- La versión original, diseñada para traducción, era más compleja.



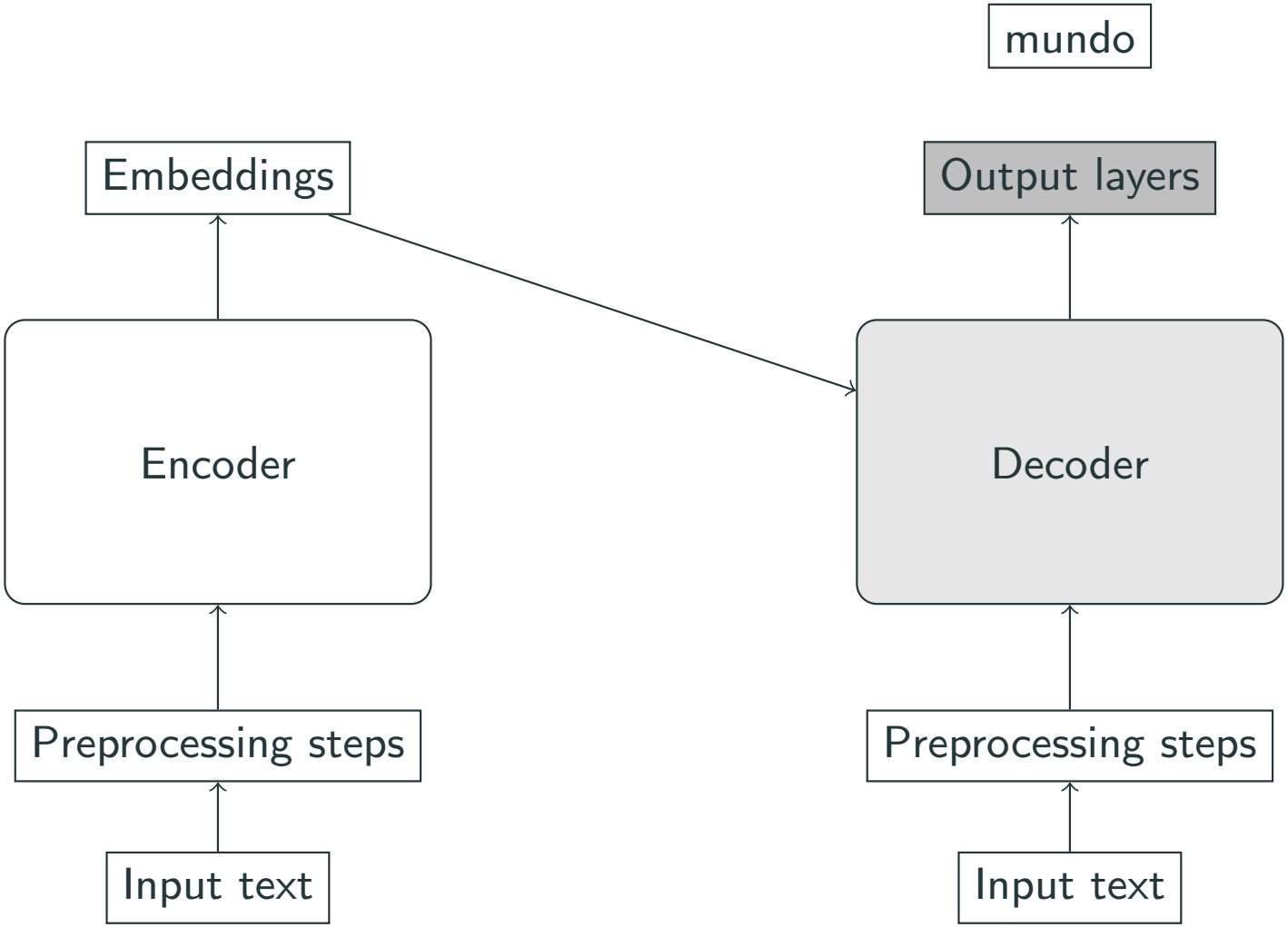
Attention is All you Need: Transformer Completo

- La versión original, diseñada para traducción, era más compleja.



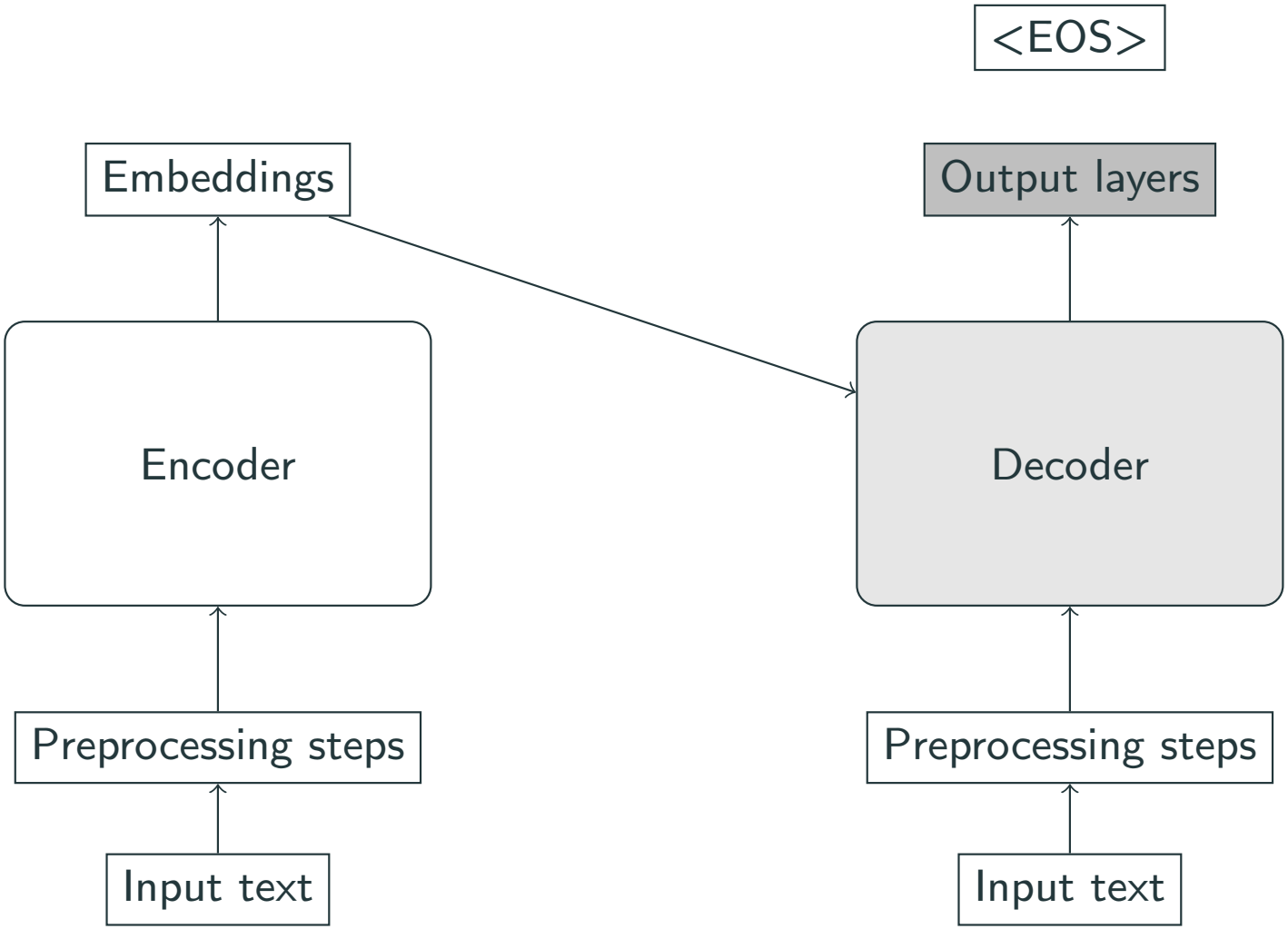
Attention is All you Need: Transformer Completo

- La versión original, diseñada para traducción, era más compleja.



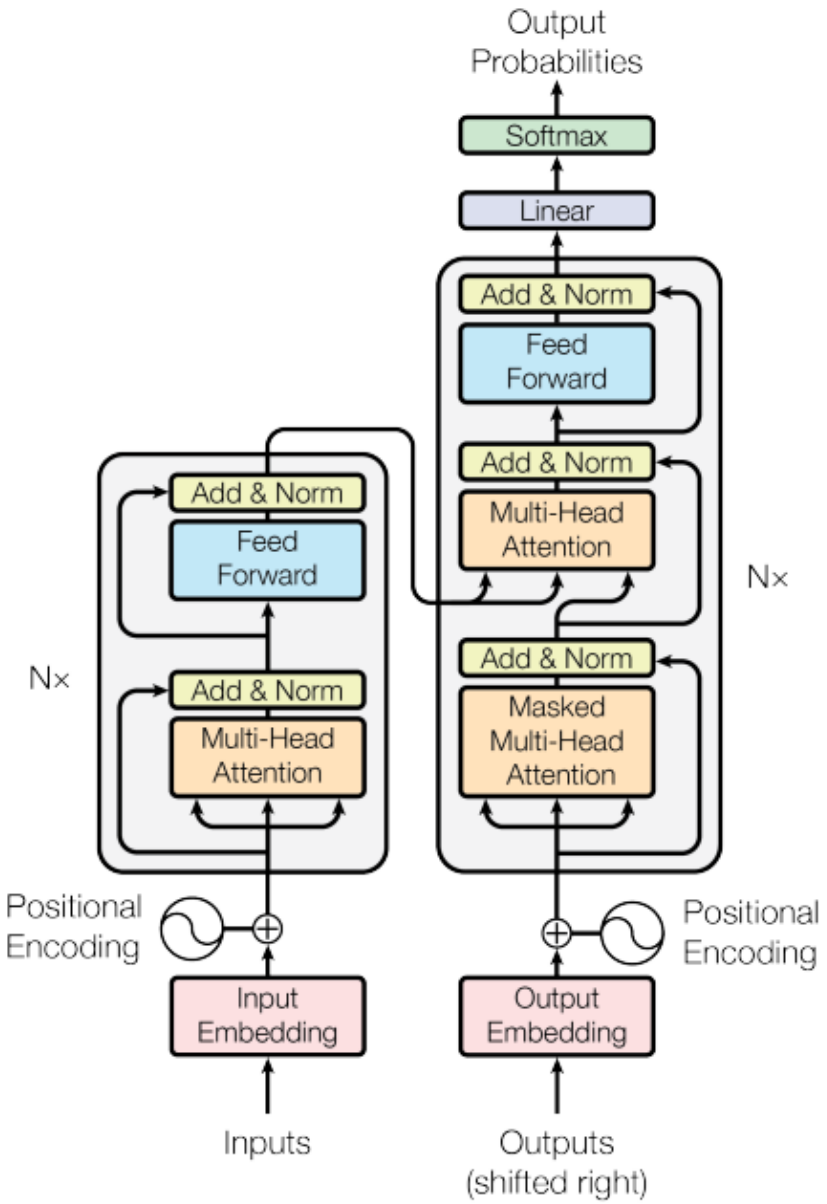
Attention is All you Need: Transformer Completo

- La versión original, diseñada para traducción, era más compleja.



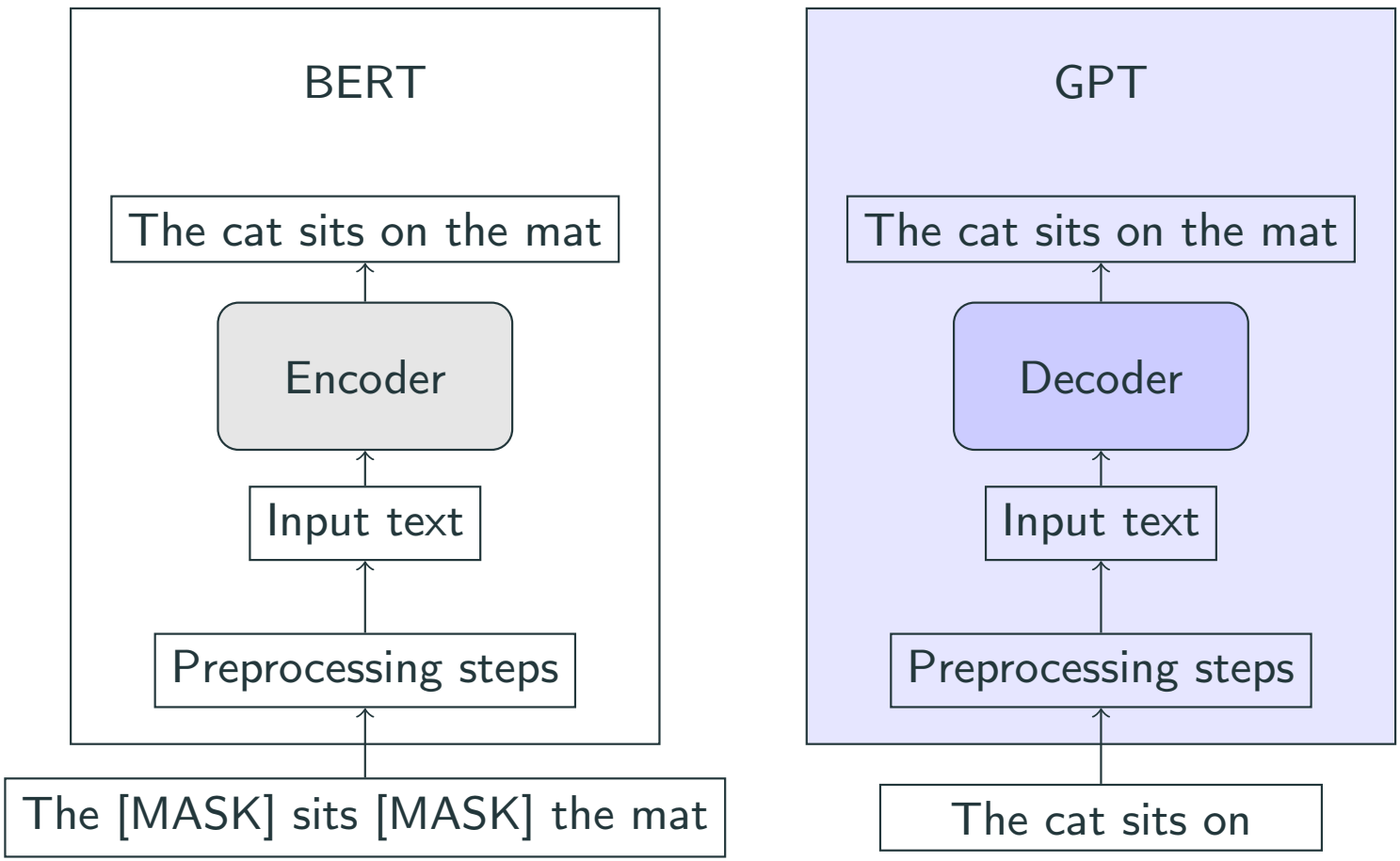
Attention is All you Need: Transformer Completo

- La versión original, diseñada para traducción, era más compleja.



Large Language Models (LLMs)

- **Large Language Models (LLMs)**: Modelos de lenguaje basados en transformers entrenados a gran escala
- Los LLMs modernos pueden basarse en arquitecturas:
  - **encoder-decoder**, como T5 (**familia T5**)
  - **encoder-only**, como BERT (**familia BERT**)
  - **decoder-only**, como GPT (**familia GPT**)



Pre-entrenamiento, Fine-tuning y Modelos Disponibles

- **Pre-entrenamiento (Pre-training)**:
  - Entrenar en **conjuntos de datos masivos** de texto (ej: Common Crawl $\approx$ 1 petabyte)
  - Objetivo: aprender representaciones generales del lenguaje
  - Costo: millones de dólares, semanas/meses de entrenamiento
- **Modelos pre-entrenados disponibles**:
  - BERT: bert-base-uncased, bert-large
  - GPT: gpt2, gpt-3.5-turbo, gpt-4
  - T5: t5-small, t5-base, t5-large
  - **Modelos distilled**: DistilBERT, TinyBERT (más pequeños y rápidos)
- **Fine-tuning** (ajuste fino):
  - Partir del modelo pre-entrenado
  - Adaptar a tarea específica con pocos datos etiquetados
  - *Veremos esto en detalle en el módulo 4*

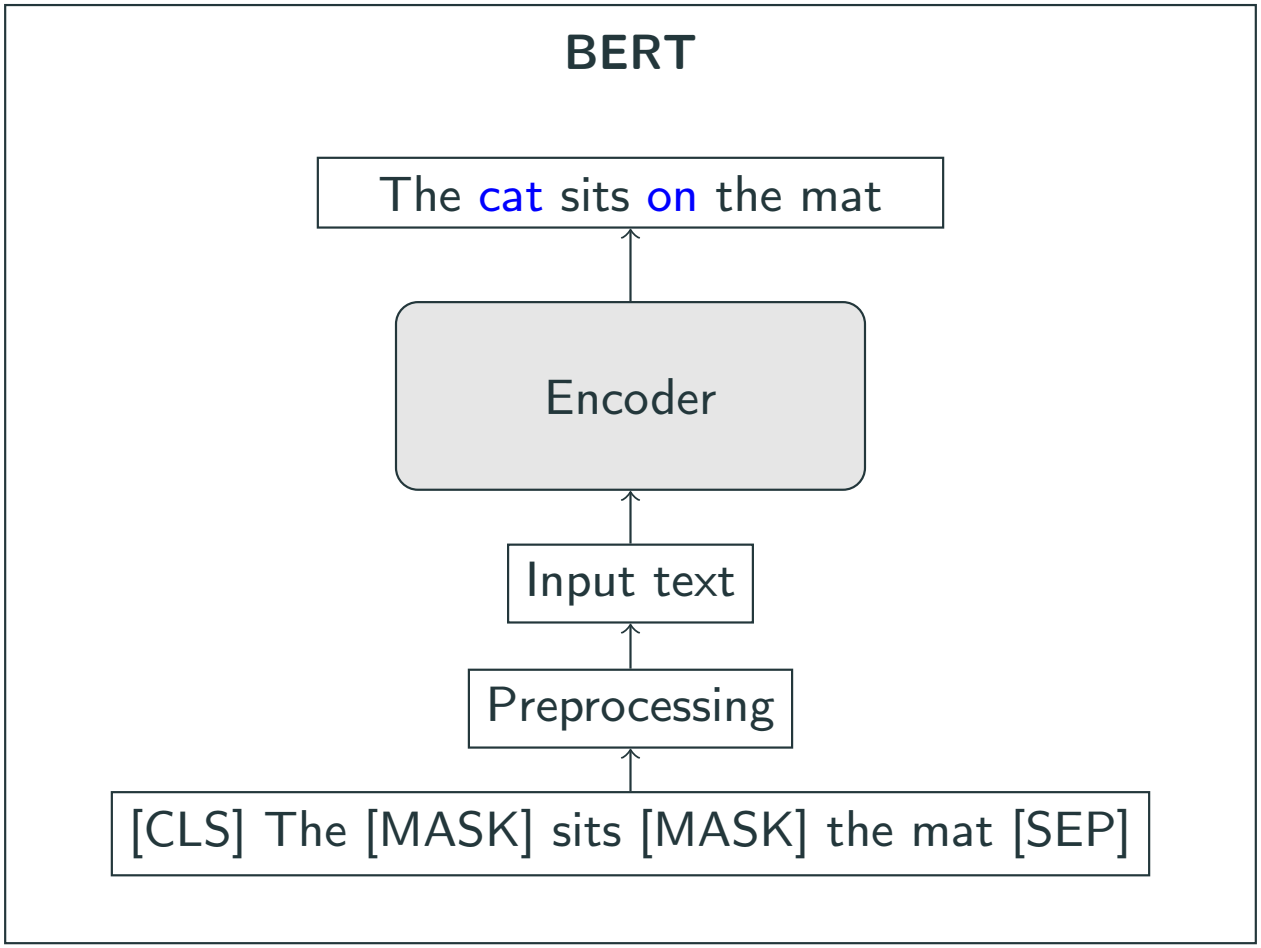
¿Cuándo usar Encoder-Decoder Vs. Encoder-only Vs. Decoder-only?

- **Encoder-Decoder** como **T5**:
  - Mejor para tareas de transformación texto-a-texto: traducción, resumen, parafraseo
  - Combina comprensión bidireccional (encoder) con generación autoregresiva (decoder)
  - Cuando necesitas tanto entender como generar texto de salida diferente al de entrada
- **Encoder-only** (como **BERT**):
  - Mejor para tareas de comprensión: clasificación, análisis de sentimiento, reconocimiento de entidades
  - Comprensión de contexto bidireccional
  - Cuando necesitas analizar o extraer información del texto
- **Decoder-only** (como **GPT**):
  - Mejor para tareas generativas: generación de texto, completado, traducción
  - Generación de texto auto-regresiva
  - Cuando necesitas generar texto coherente o continuar secuencias

Modelos BERT

Modelos BERT

- BERT significa **Bidirectional Encoder Representations from Transformers**
- Usa solo la parte **Encoder** del Transformer original



BERT: Masked Language Modeling (MLM)

- **Objetivo de pre-entrenamiento**: Predecir tokens enmascarados usando contexto bidireccional (Clasificación)
- **Formalización**:
  - Predecir tokens enmascarados: $P(x_i|\tilde{\mathrm{x}})$ donde $\tilde{\mathrm{x}}$ es la secuencia con máscaras
  - Loss de MLM: $\mathcal{L}_{MLM} = -\sum_{i \in M} \log P(x_i|\tilde{\mathrm{x}})$
  - M = conjunto de posiciones enmascaradas ($\approx 15\,\%$ de tokens)
- **Estrategia de enmascaramiento** ($15\,\%$ de tokens):
  - $80\,\%$: reemplazar con [MASK]
  - $10\,\%$: reemplazar con token aleatorio
  - $10\,\%$: mantener token original
- **Atención bidireccional**:
  - Puede atender a tokens anteriores **y posteriores**
  - Vs. GPT que solo atiende a tokens anteriores (causal)
  - Mejor para tareas de comprensión que requieren contexto completo
  - Ejemplo: “The [MASK] sits on the mat"
    - Usa “The"(izquierda) + “sits on the mat"(derecha)
    - Predice: “cat", “dog", “bird", etc.

BERT y el token [CLS]

Uso del token [CLS] para clasificación:

- El encoder procesa [CLS] con atención bidireccional sobre toda la secuencia
- La representación final de [CLS] captura información de TODO el input
- Se usa como vector de entrada para tareas de clasificación

Example: Análisis de sentimiento

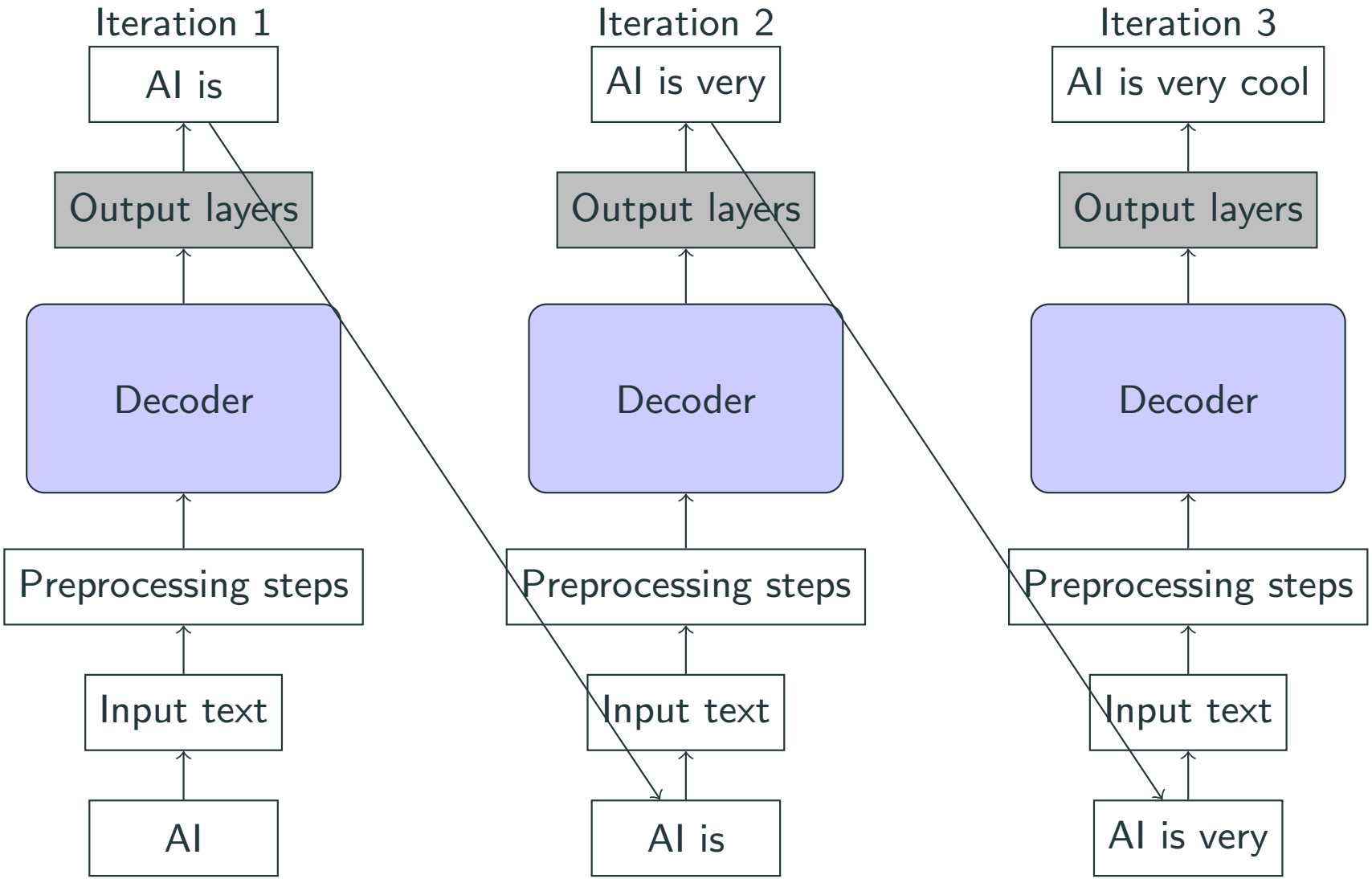
Input: “[CLS] This movie was amazing [SEP]"

[CLS] $\rightarrow$ capa de clasificación $\rightarrow$ “Positivo"

Modelos GPT

Modelos GPT

- GPT significa Generative Pre-trained Transformer.



GPT: Causal Language Modeling

- **Objetivo de pre-entrenamiento**: Predecir el siguiente token dada la secuencia anterior (Clasificación)
- **Formalización**:
  - Maximizar: $P(x_t|x_{<t}) = P(x_t|x_1, x_2, ..., x_{t-1})$
  - O equivalente: $\mathcal{L} = \sum_{t=1}^{T} \log P(x_t|x_{<t})$
- **Atención causal (unidireccional)**:
  - Solo puede atender a tokens **anteriores** (left-to-right)
  - Vs. BERT que atiende en ambas direcciones
  - Implementado con máscara triangular en self-attention
- **Ejemplo**: Generar “AI is very cool"
  - $t = 1$: predice “is"dado “AI"
  - $t = 2$: predice “very"dado “AI is"
  - $t = 3$: predice “cool"dado “AI is very"

Scale Matters: Capacidades Emergentes

Capacidades Emergentes en Modelos Grandes

- **Capacidades emergentes**: Habilidades que aparecen al escalar modelos (parámetros, datos, cómputo)
- **Ejemplos de capacidades emergentes**:
  - Razonamiento aritmético complejo (sumas de 3+ dígitos)
  - Traducción entre idiomas no vistos en entrenamiento
  - Resolución de problemas multi-paso
  - Comprensión de instrucciones complejas
- **Scaling Laws** (Kaplan et al., 2020):
  - Al aumentar escala:
    - Mejor rendimiento...
    - ... y **habilita comportamientos** ausentes en modelos pequeños.
  - Hipótesis muy debatida hoy en día. ¿Se ha alcanzado un Plateau?