Transformers II: Arquitectura y Escalado

De la atención al modelo completo

Cesar Garcia

2025

Introducción

- En la sesión anterior estudiamos **la atención** como mecanismo
- Hoy veremos cómo se **ensambla** un Transformer completo
- Introducimos cabezas múltiples, residuales y máscaras
- Conectamos arquitectura con escalabilidad

¿Por qué la atención por sí sola no define un modelo completo?

Recordatorio: atención

Atención como bloque

La atención:

- conecta todos los tokens
- produce representaciones contextualizadas
- no impone orden secuencial

Necesitamos **estructura adicional**.

Multi-Head Attention

Motivación

Una sola atención aprende un único patrón.
Multi-head attention permite:

- múltiples subespacios
- relaciones simultáneas

¿Qué ventaja tiene dividir la atención?

Multi-Head Attention

Definición formal

Dividimos el embedding en h cabezas:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

Concatenamos y proyectamos:

$$\text{MHA}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O$$

Conexiones Residuales

Motivación

Redes profundas sufren:

- degradación
- gradientes débiles

La solución:

$$X + \text{Sublayer}(X)$$

Layer Normalization

Estabilidad

Después de cada subcapa:

$$\text{LayerNorm}(X)$$

Beneficios:

- entrenamiento estable
- mejor escalado

Feedforward Network

Capa interna

Cada token pasa por una MLP independiente:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

Bloque Transformer

Estructura

Un bloque encoder:

1. Multi-head self-attention
2. Add + LayerNorm
3. Feedforward
4. Add + LayerNorm

Encoder vs Decoder

Diferencias

Encoder:

- atención completa
- ve toda la secuencia

Decoder:

- atención enmascarada
- solo ve el pasado

Máscara causal

Autoregresión

Para evitar mirar al futuro:

$$\text{mask}_{ij} = \begin{cases} 0 & j \leq i \\ -\infty & j > i \end{cases}$$

Por qué escalan los Transformers

Claves

- paralelismo total
- dependencias globales
- arquitectura homogénea

Idea clave de la sesión

Arquitectura para escalar

Los Transformers no solo modelan secuencias: escalan con datos y cómputo.