Transformer Architectures

1. Introduction

Les architectures Transformer ont révolutionné le traitement du langage naturel en abandonnant les réseaux récurrents au profit de mécanismes d'attention.

Introduites par Vaswani et al. en 2017, elles sont à la base de modèles tels que BERT, GPT ou T5.

2. Mécanisme d'attention

Le mécanisme d'attention calcule des pondérations entre chaque paire de tokens d'entrée.

Formellement, Attention(Q, K, V) = softmax(QK^T / sqrt(d_k)) V.

Cette opération permet de capturer des dépendances à longue distance.

3. Architectures multi-tête et positionnal encodings

L'attention multi-tête permet au modèle de se concentrer sur différentes sous-espaces.

Les encodages positionnels ajoutent une information de position au modèle, souvent sinusoïdale.

4. Applications et limites

Les Transformers sont utilisés pour la traduction, la génération de texte, la classification.

Limitations : coût computationnel élevé, besoin de grandes quantités de données.