L'attention dans les modèles Transformers

Résumé

Le **Transformer**, introduit par Google en 2017 dans l'article *"Attention Is All You Need"*, a révolutionné le **traitement du langage naturel (NLP)** en permettant un entraînement plus efficace des modèles de langage. Contrairement aux réseaux récurrents (RNN), le Transformer traite simultanément l'ensemble des mots d'une phrase en encodant leur position, ce qui améliore la vitesse d'apprentissage et la compréhension du contexte.

Le modèle est composé de deux principales parties :

- **L'encodeur** qui traite l'ensemble de la phrase d'entrée en appliquant une **auto-attention** (self-attention), permettant d'évaluer les relations entre les mots.
- **Le décodeur** qui génère une réponse en tenant compte des résultats de l'encodeur et de la sortie générée précédemment.

L'un des aspects clés du Transformer est le **mécanisme d'auto-attention**, qui repose sur trois vecteurs fondamentaux : **Query (Q), Key (K) et Value (V)**. Ce mécanisme permet d'accorder plus d'importance à certains mots en fonction du contexte, facilitant ainsi des tâches complexes comme la traduction automatique.

Une extension de ce concept est l'**attention multi-tête**, où plusieurs couches d'auto-attention sont utilisées simultanément pour capturer des relations complexes dans le texte. Cela améliore considérablement la qualité des résultats générés.

Les modèles basés sur cette architecture, comme **BERT, GPT et PaLM**, sont devenus incontournables dans le domaine de l'IA et du NLP. Grâce aux **techniques d'apprentissage par renforcement et de fine-tuning**, ces modèles sont continuellement optimisés pour mieux comprendre et générer du texte de manière naturelle.

Terme	Définition
Transformer	Architecture de réseau neuronal introduite en 2017, optimisée pour le NLP grâce à l'auto-attention.
Attention	Mécanisme qui permet au modèle de se concentrer sur les mots importants d'une phrase.
Auto-attention (Self-attention)	Processus qui calcule l'importance de chaque mot par rapport aux autres mots d'une phrase.
Query (Q), Key (K), Value (V)	Vecteurs utilisés dans le calcul de l'auto-attention pour déterminer la relation entre les mots.
Attention multi-tête	Technique permettant au modèle d'analyser plusieurs relations contextuelles simultanément.
Encodeur	Partie du Transformer qui analyse l'ensemble du texte d'entrée et capture les relations entre les mots.
Décodeur	Partie du Transformer qui génère du texte en fonction des données traitées par l'encodeur.
Softmax	Fonction mathématique utilisée pour normaliser les scores de l'auto-attention.
Fine-tuning	Affinage d'un modèle pré-entraîné sur une tâche spécifique pour améliorer ses performances.
RLHF (Reinforcement Learning from Human Feedback)	Technique d'apprentissage qui ajuste les modèles grâce aux retours humains.