Les grands modèles de langage (LLM) - 1

Résumé

L'essor des modèles de langage (LLM) a connu une accélération majeure avec l'introduction du **Transformers** par Google en 2017 dans l'article *"Attention Is All You Need"*. Ces modèles, comme **BERT**, **GPT**, et **ELMo**, ont révolutionné le traitement du langage naturel (NLP) en comprenant le contexte des mots d'une manière inédite.

Évolution des modèles de langage

1. **BERT (2018, Google AI)** : Premier modèle bidirectionnel, capable de comprendre le contexte d'un mot en tenant compte des mots qui le précèdent et le suivent.
2. **GPT-3 (2020, OpenAI)** : Modèle génératif avec 175 milliards de paramètres, entraîné sur un large corpus de texte. Capable de résumer, traduire, classifier du texte et générer du code.
3. **GPT-4 (2023, OpenAI)** : Évolution majeure avec une capacité contextuelle améliorée (32K tokens).
4. **Autres modèles** :
    - **T-NLG (Microsoft, 2020)** : 17 milliards de paramètres.
    - **PaLM (Google, 2022)** : 540 milliards de paramètres.
    - **LLaMA (Meta, 2023)** : Conçu pour la recherche et l'optimisation des ressources.
    - **Mixtral (Mistral AI, 2023)** : Modèle modulaire performant.

Entraînement et limitations

L'entraînement des LLM se fait en plusieurs étapes :

1. **Pré-entraînement auto-supervisé** : Apprentissage de la grammaire et du contexte à partir de grands corpus.
2. **Fine-tuning** : Adaptation du modèle pour des tâches spécifiques comme l'analyse de sentiments.
3. **RLHF (Reinforcement Learning from Human Feedback)** : Amélioration des réponses du modèle en intégrant des retours humains.

Cependant, ces modèles ont des **limites** :

- **Hallucinations** : Production d'informations erronées.
- **Biais** : Influence des données d'entraînement.
-
- **Consommation énergétique** : Coût élevé en ressources informatiques.

Perspectives et défis

L'avenir des LLM repose sur leur optimisation pour une meilleure interprétabilité, réduction des biais et efficacité énergétique. Des modèles comme **ChatGPT** et **Claude (Anthropic)** s'efforcent d'améliorer l'interaction avec les humains tout en minimisant les erreurs.

Terme	Définition
LLM (Large Language Model)	Modèle d'IA entraîné sur de grandes quantités de texte pour générer ou analyser du langage naturel.
Transformer	Architecture de réseau neuronal introduite en 2017, basée sur le mécanisme d'attention.
GPT (Generative Pre-trained Transformer)	Famille de modèles développés par OpenAI, spécialisés dans la génération de texte.
BERT (Bidirectional Encoder Representations from Transformers)	Modèle de Google permettant une compréhension du contexte en analysant les mots avant et après un terme donné.
Fine-tuning	Ajustement d'un modèle pré-entraîné sur une tâche spécifique.
RLHF (Reinforcement Learning from Human Feedback)	Méthode d'apprentissage où les retours humains guident l'amélioration du modèle.
Hallucination	Génération de réponses incorrectes ou fictives par un modèle de langage.
Token	Unité de texte utilisée par un modèle, pouvant représenter un mot, un caractère ou une sous-partie d'un mot.
Fenêtre de contexte	Nombre maximal de tokens qu'un modèle peut prendre en compte simultanément pour générer une réponse.