Nouh CHELGHAM, Fahd FADHA, Noura FAIZ, Hajar FORSI DROBI, Adil GHALEM

Multimodal Large Language Models

INTRODUCTION

Ces dernières années, les progrès en apprentissage profond et la large disponibilité de données textuelles ont favorisé l'essor des *Large Language Models* (LLM). Ces LLM sont capables de générer du texte cohérent et pertinent et sont reconnus dans le domaine du traitement automatique du langage naturel (TALN). Désormais, une nouvelle problématique concerne l'intégration efficace du langage avec d'autres modalités telles que les images, l'audio et la vidéo, donnant naissance aux *Multimodal Large Language Models* (MLLM).

LARGE LANGUAGE MODELS : LE CERVEAU

Pourquoi on est passé aux LLMs? Et c'est quoi un LLM?

Les chercheurs ont constaté que les LM de grande taille améliorent les performances et révèlent des capacités exceptionnelles, telles que l'apprentissage selon le contexte.

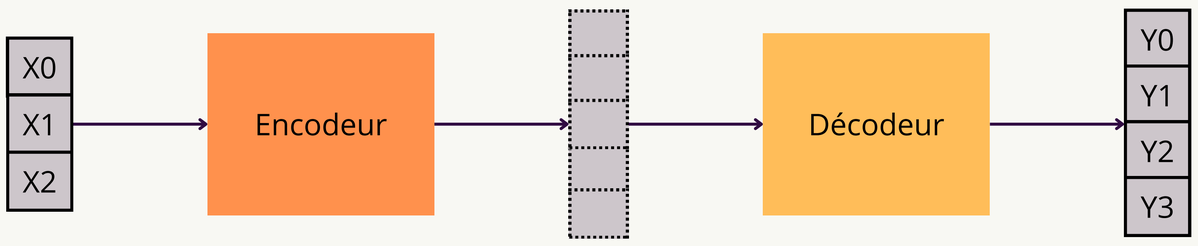
Entraînés sur d'immenses ensembles de données textuelles et possédant des milliards de paramètres, ces LLMs démontrent une compréhension remarquable du langage naturel et sont capables d'exécuter diverses tâches. Ils reposent souvent sur l'architecture *Transformer*, reconnue pour sa facilité de parallélisation, ce qui leur permet d'atteindre des tailles sans précédent.

TRANSFORMER

Encodeur : Utilise des blocs d'attention multi-têtes pour calculer l'importance des mots dans une séquence, tandis que le

Décodeur : Prédit la séquence de sortie en se basant sur l'information de l'encodeur et sur la séquence cible.

Capture dépendances à long terme



L'attention

Déterminer quels éléments de l'entrée sont les plus pertinents pour la tâche en cours. Peut aider le modèle à mieux comprendre le contexte et à produire des résultats plus précis.

L'ÉVOLUTION DES *LANGUAGE MODELS*

Statistical Language Models (SLM)

Les SLM prédisent le mot suivant en se basant sur le contexte récent, mais rencontrent des difficultés avec les dépendances à long terme, les nuances sémantiques et les mots hors vocabulaire.

Neural Language Models (NLM)

Les NLM utilisent des architectures de réseaux neuronaux, et excellent dans la capture de modèles linguistiques et de dépendances grâce à l'utilisation de représentations de mots distribuées.

Pre-trained Language Models (PLM)

Les PLM révolutionnent le TALN en entraînant des modèles de langage bidirectionnels pour la représentation des mots en fonction du contexte, introduisant des paradigmes de pré-entraînement et de réglage fin qui améliorent considérablement l'efficacité des modèles.

Représentation distribuée des mots

Consiste à encoder chaque mot sous forme d'un vecteur de nombres réels dans un espace multidimensionnel, capturant ainsi les relations sémantiques et syntaxiques entre les mots. Ces représentations sont apprises à partir de grands ensembles de données textuelles à l'aide de techniques d'apprentissage automatique, et permettent d'améliorer la compréhension et les performances des LLM.

MULTIMODAL LARGE LANGAGE MODELS (MLLM)

Encoder and Joint Embedding Space espace d'encodage et d'incorporation conjoints permettant de représenter de manière cohérente et unifiée des données provenant de différentes modalités, telles que le texte, l'image et l'audio

Natural language supervision paradigme d'apprentissage où les modèles sont entraînés en utilisant des données supervisées dans un langage naturel, permettant une compréhension et une génération plus précises du langage humain

Applications

- Description d'images
- Réponse à des questions visuelles
- Génération d'images à partir de texte

Contrastive leanning maximiser la similarité entre des paires d'échantillons positifs tout en minimisant la similarité entre des paires d'échantillons négatifs dans un espace d'incorporation commun

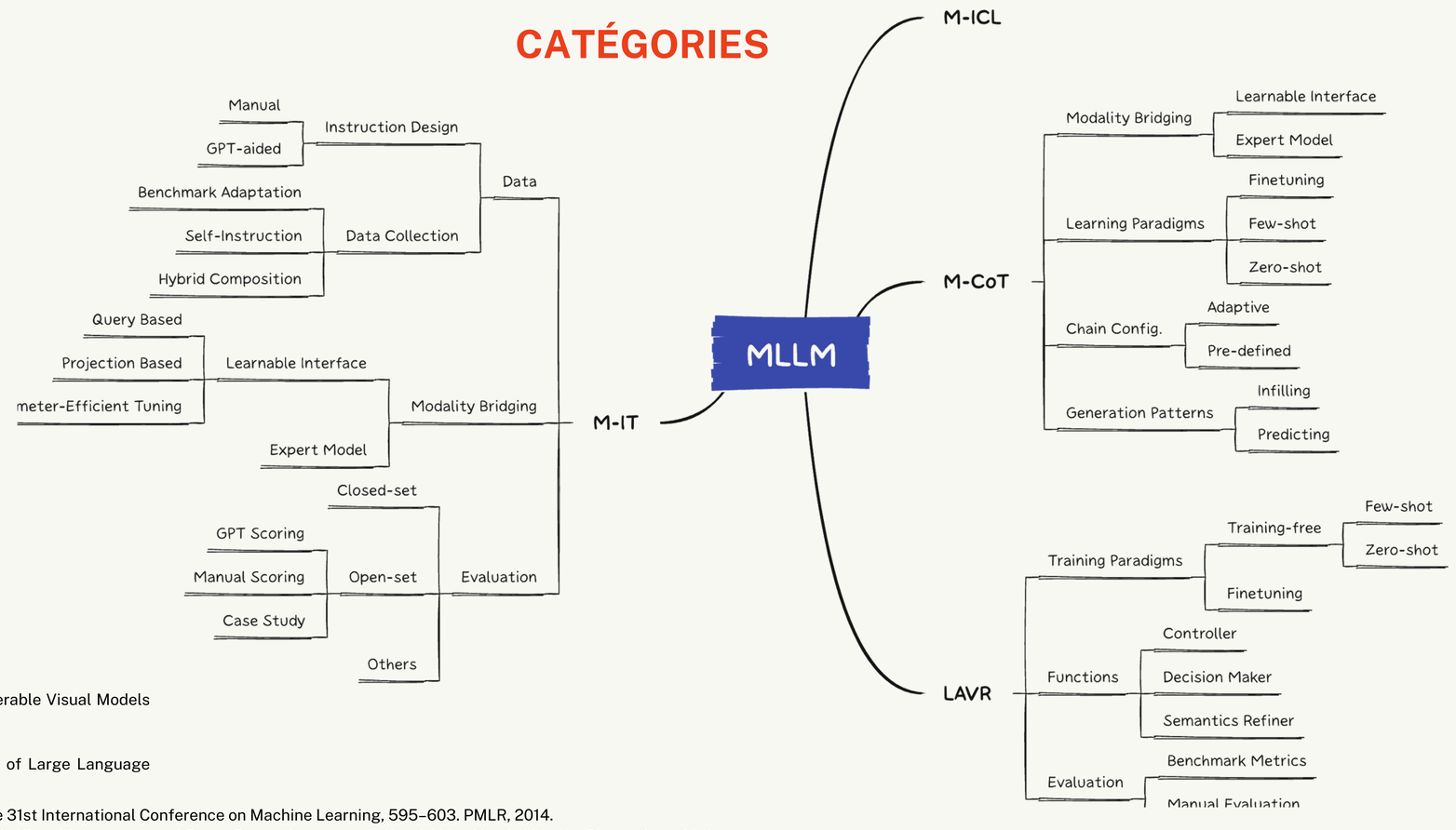
DISCUSSION

Les modèles de langage large (LLM) et multimodaux (MLLM) démontrent des capacités remarquables en compréhension linguistique et en exécution de diverses tâches, mais leur objectif souvent ambigu ou vaste rend difficile la conduite de tests approfondis.

L'optimisation des modèles axée sur l'exactitude a conduit à une forte dépendance envers des ressources informatiques étendues, posant des défis environnementaux et d'efficacité des ressources.

Naviguer dans les tâches multimodales exige une compréhension profonde des interactions entre différents types de données pour améliorer la fiabilité et les performances des systèmes.

CATÉGORIES



RÉFÉRENCES

1] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... and Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. arXiv preprint arXiv:2103.00020.
\newblock {\em ACM Computing Surveys (CSUR)}, 56(2):Article 30, 2023.
[2]Zhao, Wayne Xin, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, et al. "A Survey of Large Language Models." arXiv, November 24, 2023.
[3]Kiros, Ryan, Ruslan Salakhutdinov, and Rich Zemel. "Multimodal Neural Language Models." In Proceedings of the 31st International Conference on Machine Learning, 595–603. PMLR, 2014.
[4] Huang, Shaohan, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, and Barun Patra. "Language Is Not All You Need: Aligning Perception with Language Models." Advances in Neural Information Processing Systems 36 (2024).
[5]Shengbang Tong and Erik Jones and Jacob Steinhardt , 2023 , "Mass-producing Failures of Multimodal Systems with Language Models" , 2306.12105.
[6]Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth.
\newblock "Recent Advances in Natural Language Processing via Large Pre-trained Language Models: A Survey."
[7] Kiros, Ryan, Ruslan Salakhutdinov, and Rich Zemel. "Multimodal Neural Language Models." In Proceedings of the 31st International Conference on Machine Learning, 595–603. PMLR, 2014.
[8] Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention." In Proceedings of the 32nd International Conference on Machine Learning, 2048–57. PMLR, 2015.
[9] Yan Zeng , Hanbo Zhang , Jiani Zheng ,Jiangnan Xia, Guoqiang Wei, Yang Wei, Yuchen Zhang, Tao Kong . "What Matters in Training a GPT4-Style Language Model with Multimodal Inputs?".
[10] Wu, Carole-Jean, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, et al. "Sustainable AI: Environmental Implications, Challenges and Opportunities." Proceedings of Machine Learning and Systems 4 (April 22, 2022): 795–813.
[11]Lu, J., Batra, D., Parikh, D., and Lee, S. (2019). Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2137-2146).