Laboratoire de l'Informatique du Parallélisme
(UMR CNRS - ENS Lyon - UCB Lyon 1 - INRIA 5668)
équipe ROMA

sujet de stage

Allocation optimisée d'experts pour l'inférence des grands modèles de langages

Loris Marchal & Olivier Beaumont

Encadrants : Loris Marchal (directeur de recherche, CNRS, `Loris.Marchal@ens-lyon.fr`) Olivier Beaumont (directeur de recherche, Inria Bordeaux, `Olivier.Beaumont@inria.fr`)

Lieu : Laboratoire LIP, École Normale Supérieure de Lyon, équipe ROMA.

Équipe d'accueil

La stage se déroulera dans l'équipe ROMA, du Laboratoire de l'Informatique du Parallélisme (UMR CNRS - ENS Lyon - UCB Lyon 1 - INRIA 5668), sous la direction de Loris Marchal. L'équipe ROMA s'intéresse à la conception d'algorithmes parallèles et d'ordonnancements pour les plates-formes de calcul distribuées, en particulier pour les applications scientifiques. Le stage sera également encadré par Olivier Beaumont, de l'équipe Inria Topal à Bordeaux. Topal développe des algorithmes et des outils efficaces pour l'algèbre linéaire et tensorielle et pour l'apprentissage automatique.

Sujet de stage

La génération de texte avec des grands modèles de langages (ou LLM pour *Large Languages Models*) requiert de grandes capacités de calcul et de mémoire. Le paradigme du mélange d'experts (*Mixture of Experts*, ou MoE) a été proposé pour réduire la taille des données nécessaires à l'inférence [3] : chaque couche du modèle est composée de plusieurs experts et seulement un petit nombre d'entre eux est utilisé pour la production d'un mot. Le modèle DeepSeek comporte ainsi 64 experts à chaque couche, dont seulement 8 sont actifs pour produire un mot [2].

Lors de l'inférence sur une plate-forme distribuée consitutée de plusieurs GPUs, les experts peuvent être répartis entre les GPUs, et éventuellement répliqués. On cherche en général à traiter plusieurs requêtes de générations de texte (ou *prompts*) en parallèle pour améliorer l'efficacité du calcul, cependant ce nombre de requêtes est limitée par la mémoire nécessaire au stockage de leur contexte.

L'objectif de ce stage est d'optimiser la distribution des experts sur les GPUs, leur éventuelle réplication ainsi que l'allocation des experts nécessaires pour les différentes requêtes sur les GPUs, pour améliorer le débit de génération. Dans une première étude [1], nous avons montré qu'utiliser la fréquence d'utilisation simultanée des sous-ensembles d'experts permet d'augmenter le niveau de parallélisme pour une seule requête, et donc le débit en inférence. L'objectif de ce stage est d'étendre cette étude lorsque plusieurs requêtes sont traitées simultanément, ce qui constitue à la fois un cas d'usage plus réaliste et permet d'augmenter les performances en inférence. Il s'agira de dégager des compromis entre débit d'inférence et latence pour chacune de ces requêtes.

Le travail consistera à :

- Modéliser précisément le problème d'optimisation lié à la distribution des experts et l'allocation des requêtes, ainsi qu'effectuer un travail bibliographique sur les stratégies existantes ;
- Proposer des méthodes efficaces, éventuellement avec des garanties de performances, en commençant par des cas simples (une seule requête et/ou pas de réplication) puis en les généralisant ;
- Tester les solutions proposées en simulation, sur des traces d'activation d'experts obtenues en exécutant les modèles correspondants sur des jeux de données existants.
- Éventuellement implanter certaines de ces stratégies dans un système d'inférence existant.

Compétences requises

Le stagiaire devra posséder de bonnes compétences en algorithmique et en programmation. Une compréhension de l'apprentissage automatique, des modèles de langages et une maitrise des outils classiques de leur implémentation seront des atouts importants. Une connaissance préliminaire de la recherche opérationnelle et/ou du calcul haute-performance sera également appréciée.

Collaborations et perspectives

Le stage se déroulera à Lyon, au LIP, mais sera co-encadré par Olivier Beaumont (Inria Bordeaux). Les échanges avec Olivier Beaumont se feront d'abord en visio, puis des visites pourront être organisées pour travailler ensemble. En fonction du déroulement du stage, une poursuite en thèse sera envisagée.

Nous travaillons sur ces sujets en collaboration avec Oana Balmau et Mark Coates de l'université McGill à Montréal (Canada). Si le stagiaire poursuit par une thèse, celle-ci pourra être effectuée en co-tutelle avec l'université McGill et l'ENS de Lyon.

Bibliographie

[1] Olivier Beaumont, Raphaël Bourgouin, Maxime Darrin, Loris Marchal, and Pablo Piantanida. Leveraging expert usage to speed up LLM inference with expert parallelism. In *31st European Conference on Parallel and Distributed Processing (Euro-Par)*, volume 15900 of *Lecture Notes in Computer Science*, pages 145–158. Springer, 2025.

[2] Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y Wu, et al. Deepseekmoe : Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv :2401.06066*, 2024.

[3] Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, et al. Glam : Efficient scaling of language models with mixture-of-experts. In *ICML 2022*, 2022.