LLM à jour pour tous

LLM4ALL

Programme : **TSIA**

Édition : **2023**

Contact : **cerisara@loria.fr**

COORDINATEUR :

Christophe Cerisara

PARTENAIRES :

LORIA, LIX, AP-HP
LINAGORA
Support de Huggingface

Résumé (3 lignes max) :

Ré-entraîner un LLM from scratch coûte trop cher; continuer l'apprentissage est difficile (convergence, oubli, performances). LLM4ALL vise à mettre à jour des LLM à faible coût à partir de LLM open-source, et appliquera ces méthodes au résumé de réunions et aux appels d'urgence du SAMU.

CONTEXTE ET OBJECTIFS

Comment mettre à jour un LLM ? Le finetuning ajoute difficilement de nouvelles connaissances et mène au *catastrophic forgetting*.

Objectifs:
- mettre à jour les modèles existants
- réduire les coûts d'apprentissage
- faciliter l'intégration de nouvelles informations
- appliquer au dialogue en santé et en réunion

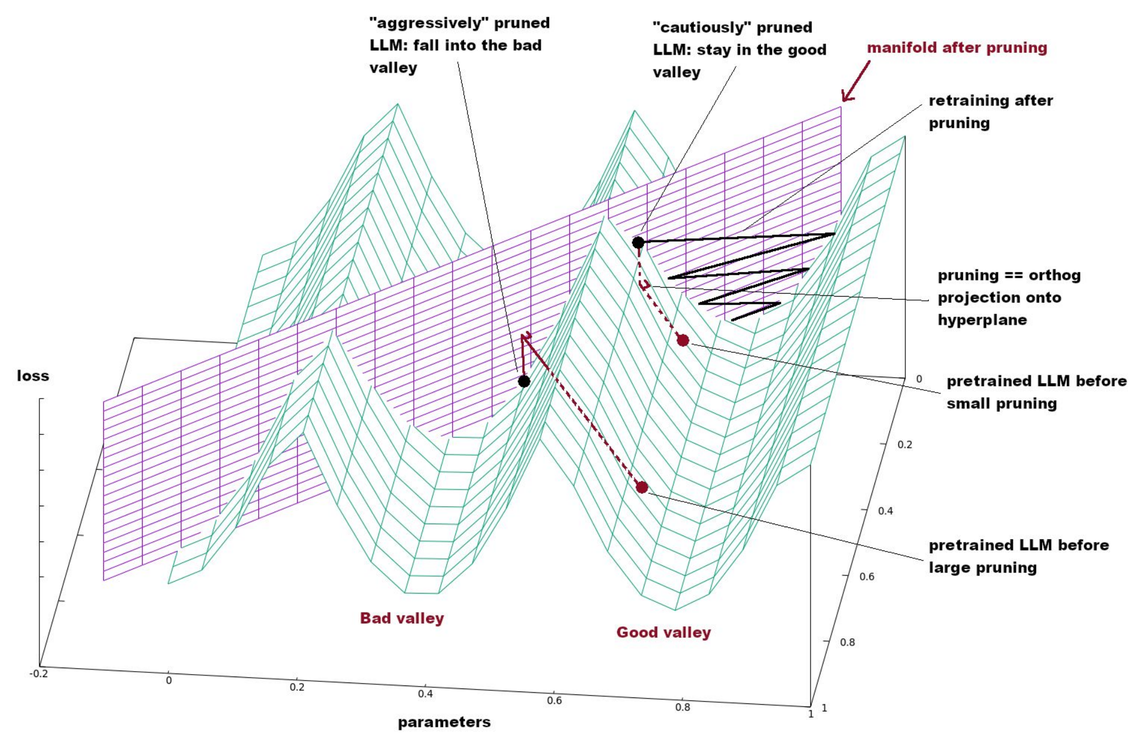

Figure: Interprétation géométrique du pruning

- Pruning = projection sur hyperplan H
- Si H est "loin", modèle projeté n'est pas *linearly connected* au modèle d'origine
- Donc il faut *iterative pruning + retraining*
- Confirmé par la *Lottery Ticket Hypothesis*

MÉTHODOLOGIE

Verrous ➪ solutions:
- Convergence trop lente avec continued pretraining ➪ 1) prune le modèle, 2) retrain [ShearedLlama]
- Coût élevé d'apprentissage ➪ Prune, compress, distill
- Catastrophic forgetting ➪ modèle grossissant, replay

LLM4ALL est lié au projet et à la communauté open-source **OpenLLM-France**

Premier modèle finetuned: Claire-Mistral-7b
- Finetuned on French Dialogue dataset (2023) ~200Mw
- 4096 context length
- Trained on Jean Zay (8x A100 for 50 gpu.h)
- User pref eval: 64% preference vs. Falcon & Mistral
- https://huggingface.co/OpenLLM-France

Highlight: compression de LLM

- Poids sont high-rank, activations sont low-rank
- Compresser activations demande plus de données
- Donc compresse les poids avec SVD puis distill pour compresser les activations

Highlight: dialogues dans les appels d'urgence

- simsamu: simulation d'appels d'urgence en FR
- 3h15 de données audio + texte, 61 session
- https://huggingface.co/datasets/medkit/simsamu

VALORISATION ET RETOMBEES

LLM4ALL produira en open-source:
- une version mise à jour d'un LLM fondation existant.
- un LLM adapté aux résumés de réunion
- un LLM adapté aux appels d'urgence du SAMU
- les études et scripts permettant de mieux comprendre les verrous et solutions pour construire ces LLM