



Optimization des Hyperparamètres appliquée au Fine Tuning de LLM

Basé sur l'article : Bayesian and Partition-Based Optimization for Hyperparameter Optimization of LLM Fine-Tuning

Nathan Dayouse

Semester A24 Soutenance ST30





Sommaire

- 1. Introduction
- 2. Design et Implémentation
- 3. Résultats et Analysis

4. Conclusion





Large Language Models

Point clés

19/02/2025

- ► Etat de l'art pour le traitement de language naturel.
- Réseaux de Neurones avec une architecture basé sur le transformer^a (annexe 1)
- ➤ Taille : entre 1 et 405 Milliards de neurones

Auto-attention



Figure: Illustration du mécanisme d'auto-attention

L'auto-attention est la clé du LLM, en permettant de comprendre le contexte

^aVaswani et al, Attention is all you need,2017





Fine Tuning

19/02/2025

Aspect	Pre-entrainement	Fine Tuning		
Objectif	Apprentissage general	Adaptation à un domaine		
Données	Larges et diverses	Restreintes et Spécifiques		
Ressources	Centaines de GPU	au moins 1 GPU		
Durée	Semaine/Mois	Heures/Jours		

Table: Comparaison entre le Pre-entrainement et le Fine Tuning de LLM

Parameter-Efficient Fine-Tuning (PEFT)

- ► Ensemble de méthodes pour réduire le nombre de paramètres à entrainer
- ► Utilisation de la méthode LoRA (annexe 2)
- ► Amène des nouveaux hyperparamètres





Optimisation des Hyperparamètres (OHP)

Hyperparamètres

Paramètres qui ne sont pas entrainés par le modèle (learning rate, dropout ...)

Objectifs

- Meilleur performance qu'en manuel
- Retirer le besoin d'expertise

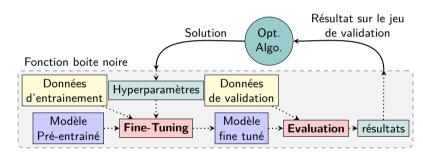


Figure: Fonctionnement général de l'optimisation des hyperparamètres

Design et Implémentation





Formulation du problème

Equation

$$\eta^* \in \arg\max_{\eta \in \mathcal{A}} f(\eta), \quad f : \mathbb{R}^d \to \mathbb{R}$$
(1)

Avec η une solution de dimension d et f la fonction représentant l'entrainement et l'évaluation d'un modèle.

Charactéristiques de la fonction f

- ► Boite-noire : non dérivable
- ► Couteux : une évaluation se compte en dizaine de minutes
- ▶ Bruité : évaluer 2 fois la même solution peut donner un résultat différent
- ► Variables mixes : les variables sont de plusieurs type (entier, continu...)





Travaux connexes

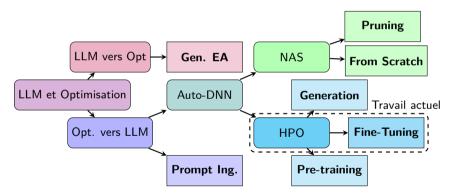


Figure: Classification des travaux similaires





Sommaire

- 2. Design et Implémentation

19/02/2025







Espace de Recherche

Hyperparamètres	Plage d'Optimisation		Туре	Conversion
пуреграгателев	Borne Inf.	Borne Sup.	туре	Conversion
Learning Rate	-10	-1	log.	$f(x)=10^x$
LoRA rank	1	64	ent.	f(x) = round(x)
LoRA scale	1	64	ent.	f(x) = round(x)
Dropout	0	0.5	cont.	f(x) = x
Weight Decay	-3	-1	log.	$f(x) = 10^x$

Table: Résumé de l'espace de recherche

► Variables mixes : étape de conversion nécessaire



19/02/2025



Search Strategy: BO

10





Search Strategy: SOO



19/02/2025



Search Strategy: BaMSOO

12





Performance Estimation Strategy





Implémentation





Sommaire

- 1. Introduction
- 2. Design et Implémentation
- 3. Résultats et Analysis
- 4. Conclusion





Expérimentation





LHS: Résultats





Résultats des 3 algorithms





Analyse





Prospectives





Sommaire

- 1. Introduction
- 2. Design et Implémentation
- 3. Résultats et Analysis

4. Conclusion





Conclusion

Une conclusion





Merci.





Annexes 1: Architecture d'un LLM

MHA, Transformers

19/02/2025





Annexes 2: Low Rank Adaptation (LoRA)

19/02/2025 _{○●} 2