Group 4

RAG vs Fine-tuning

TAN Victor & KWA MOUTOME Noah

LLMs

Sommaire

- Introduction
- Article original
- Remarques et limitations
- Nos expérimentations
- Conclusion

Introduction

RAG

Technique combinant la recherche d'information (retrieval) et la génération de texte (generation) pour améliorer les réponses d'un modèle de langage.

- Pas d'entraînement supplémentaire
- Ajout d'information externe

Fine-tuning

Processus consistant à réentraîner un modèle pré-entraîné sur un ensemble de données spécifique pour l'adapter à une tâche particulière.

- Spécialisation
- Amélioration des réponses pour la tâche

Article original

Investigating the performance of Retrieval-Augmented Generation and fine-tuning for the development of AI-driven knowledge-based systems

- But: Comparer les performances de l'ajout d'un RAG et l'utilisation de fine-tuning
- Datasets: Publications sur Urban Monitoring, Maïs, COVID
- Models: GPT-J-6B, OPT-6.7B, LLaMA-7B, LLaMA2-7B
- Métriques: ROUGE, BLEU, METEOR, Cosine Similarity

Investigating the performance of Retrieval-Augmented Generation and fine-tuning for the development of AI-driven knowledge-based systems

Róbert Lakatos[2,3,4], Péter Pollner[1], András Hajdu[2], and Tamás Joó[1,4]

[1]Data-Driven Health Division of National Laboratory for Health Security, Health Services Management Training Centre, Semmelweis University

[2]Department of Data Science and Visualization, Faculty of Informatics, University of Debrecen

[3]Doctoral School of Informatics, University of Debrecen

[4]Neumann Technology Platform, Neumann Nonprofit Ltd.

Abstract

The development of generative large language models (G-LLM) opened up new opportunities for the development of new types of knowledge-based systems similar to ChatGPT, Bing, or Gemini. Fine-tuning (FN) and Retrieval-Augmented Generation (RAG) are the techniques that can be used to implement domain adaptation for the development of G-LLM-based knowledge systems. In our study, using ROUGE, BLEU, METEOR scores,

Article original

Table 4: Average scores of each approach.

Models	ROUGE	METEOR	BLEU	CS
Baseline	0.142117	0.119251	0.002770	0.335299
Fine-tuned	0.254003	0.242348	0.050048	0.356439
RAG with fine-tuned	0.229296	0.195219	0.029378	0.305797
RAG	0.294986	0.222193	0.057998	0.544829

Résultats de l'article

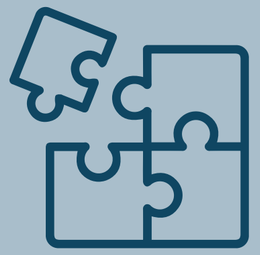
Remarques et limitations

Métriques

- BLEU et ROUGE se basent sur la correspondance des n-grams

→ Scores proches de 0 même si on a le bon sens (synonymes, …)

- METEOR et CS **peuvent corriger** ces limites



Imprécisions

- Pas d'information sur le prompt donné au LLMs

→ Alpaca model, mais quel texte ? Défaut ? Améliorable ?

- Pas d'information sur le nombre n pour le choix des n-grams

Nos expérimentations

Fine-tuning avec Unsloth

Modèles llama 1B, 3B, 8B

QLoRA avec rank 2, 4, 8, 16

Embedder all-MiniLM-L6-v2

Répétition des expériences (moyennes)

Modèle

- llama 3.2 (3B)

Dataset

- Médical
- Très spécifique / spécialisé: Jeu vidéo

Métriques

- BLEU-4, ROUGE-1, (ROUGE-2), ROUGE-L, METEOR, CS, BERTSCORE (F1)

Ajouts

- Temps de fine-tuning
- Temps d’inférence (avec ou sans RAG)

Datasets

Médical

Dataset Hugging Face

10 000 paires Question/Réponse

90% train, 10% test

Exemple de paires:

- What conditions are associated with low Ca2+ and low PTH?
- Low Ca2+ and low PTH is seen in primary hypoparathyroidism

Jeu Vidéo - League of Legends

Dataset homemade League of Legends

5 000 paires Question/Réponse

Questions générales ou plus spécifiques

Exemple de paires:

- What is the cooldown of the Q spell of Mel at level 5?
- Cooldown of Mel's Q at level 5: 6

Exemple - Evaluation

prompt = """You are a physician. Below is a question. Write a response that appropriately answers the question.

Question:
'Which drug is used for the prevention of cluster headaches?'

Answer:
{}"""

GT:
'Verapamil is the drug used for prophylaxis of cluster headaches.'

Generation:
'Sumatriptan is the drug used for the prevention of cluster headaches.'

Exemple - RAG

prompt = """"You are a physician. Below is a question. You have access to a database, where you retrieved some potential information about the question. Write a response that appropriately completes the request. If you don't know the answer, just say that you don't know. If possible, use the database and find an answer as close as the one in the database.

Question:
'Which drug is used for the prevention of cluster headaches?'

Database:
{top.1: 'Verapamil is the drug used for prophylaxis of cluster headaches.'
top.2: ....}

GT:
'Verapamil is the drug used for prophylaxis of cluster headaches.'

Answer:
{}""""

Generation:
'Verapamil is the drug used for prophylaxis of cluster headaches.'

Résultats - Médical

Type	BLEU-4	ROUGE-1	ROUGE-L	METEOR	BERT	CS	Temps
Base	0.044	0.126	0.105	0.123	0.632	0.247	7:10
Fine-tuned	0.0518	0.248	0.211	0.203	0.846	0.443	(44:03) + 24:18
RAG	0.399	0.476	0.443	0.475	0.900	0.705	48:05
Fine-tuned + RAG	0.255	0.578	0.535	0.543	0.931	0.870	(44:03) + 40:54

Résultats - League of Legends

Type	BLEU-4	ROUGE-1	ROUGE-L	METEOR	BERT	CS	Temps
Base	0.004	0.128	0.112	0.158	0.500	0.359	7:45
Fine-tuned	0.045	0.445	0.388	0.462	0.864	0.678	(09:50) + 24:15
RAG	0.174	0.317	0.302	0.360	0.859	0.520	30.0
Fine-tuned + RAG	0.431	0.683	0.675	0.687	0.930	0.770	(9:50) + 37:27

Remarques

- Rank QLoRA: 2~4 meilleur que 8~16 (5~10%)

- Rang 4 sur modèle 3B: 6,078,464/3,000,000,000 paramètres entraînés (0.20%) → 44 min (médical)
- Rang 8 sur modèle 8B: 20,971,520 / 8,000,000,000 paramètres entraînés (0.26%) → 2h30 (médical)

- Utiliser un mauvais prompt peut réduire considérablement les résultats

- Matériel: Service cloud (Nvidia A2 16GB)

Conclusion

- Contrairement à l'article, RAG + Fine-tuning **peut** être "meilleur" selon l'utilisation (métrique et prompt)
- Utiliser plus de métriques semble intéressant
- Utiliser des données totalement inconnues au modèle est intéressant

- Par nature, RAG est moins sujet aux hallucinations par rapport au Fine-tuning
- Par nature, Fine-tuning pourra mieux généraliser que le RAG

RAG ou Fine-Tune + RAG ?

- Médical: Fine-tune + RAG donne +3% à +15% de meilleurs scores (sauf BLEU) pour +77% de temps
- Jeu: Fine-tune + RAG donne +8% à +145% de meilleurs scores pour +55% de temps

Autres pistes

- Plusieurs domaines différents → MoE
- Comparer Modèle 3B vs MoE 3x1B
- Exemples de situtations où le Fine-tuning serait meilleur que RAG (généralisation, ...)
- D'autres métriques qui prennent en compte les reformulations

Merci

Références:

Article originel: Investigating the performance of Retrieval-Augmented Generation and fine-tuning for the development of AI-driven knowledge-based systems, Robert Lakatos, Peter Pollner, Andras Hajdu, Tamas Joo

Données: medalpaca/medical_meadow_medical_flashcards, sur HuggingFace
https://wiki.leagueoflegends.com/en-us/

Entraînement: https://unsloth.ai/