ELAD: Explanation-Guided Large Language Models Active Distillation

Yifei Zhang, Bo Pan, Chen Ling, Yuntong Hu, Liang Zhao
Emory University

ACL 2024 Bangkok, Thailand

Abstract

The deployment and application of Large Language Models (LLMs) is hindered by their memory inefficiency, computational demands, and the high costs of API inferences. Traditional distillation methods, which transfer the capabilities of LLMs to smaller models, often fail to determine whether the knowledge has been sufficiently transferred, potentially resulting in high costs or incomplete distillation. In this paper, we propose an Explanation-Guided LLMs Active Distillation (ELAD) framework that employs an active learning strategy to optimize the balance between annotation costs and model performance. To improve efficient sample selection, we introduce an explanation-guided sample selection method that identifies samples challenging its reasoning by exploiting uncertainties in explanation steps. Additionally, we present a customized LLM-annotated explanation revision technique where the teacher model detects and corrects flaws in the student model's reasoning. Our experiments across various reasoning datasets demonstrate that our framework significantly enhances the efficiency of LLM knowledge distillation.

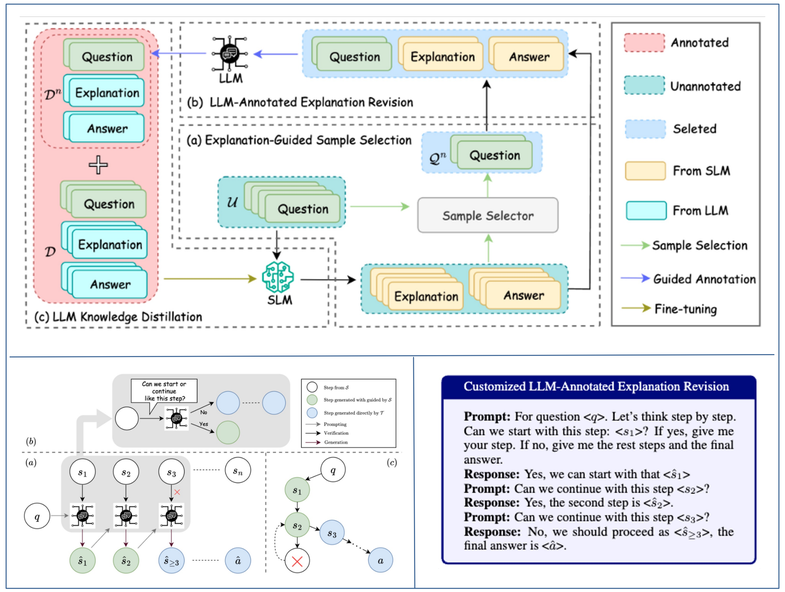
Introduction

The Introduction section of the ELAD paper discusses the significant challenges in deploying Large Language Models (LLMs) due to their substantial computational demands and high costs associated with API inferences. Traditional distillation methods, which aim to transfer the capabilities of LLMs to smaller models, often do not ensure that essential knowledge is effectively transferred, leading to high costs or inadequate performance.

To tackle these issues, the paper introduces the Explanation-Guided LLMs Active Distillation (ELAD) framework. This framework optimizes the balance between annotation costs and model performance through an active learning strategy. It features innovative methods like explanation-guided sample selection and customized explanation revision, enhancing the efficiency and effectiveness of the distillation process. These techniques not only improve the selection of training samples based on explanation uncertainties but also refine the student model's reasoning with corrections from a teacher model. The ELAD framework aims to maintain the sophisticated reasoning capabilities of LLMs in smaller models, thus reducing costs while preserving performance, setting the stage for detailed exploration of these methods in subsequent sections of the paper.

Methods and Materials

The ELAD framework employs a sophisticated two-method approach to enhance the distillation of Large Language Models (LLMs) into more efficient student models through active learning: Explanation-Guided Sample Selection (EGSS): This technique advances traditional sample selection by utilizing the explanations generated by LLMs to identify samples that reveal uncertainties in model reasoning. By focusing on samples that challenge the student model's reasoning abilities, EGSS ensures that the training process is both targeted and highly informative, promoting a deeper understanding of complex problem-solving. Customized LLM-Annotated Explanation Revision (CLEAR): In this step, a teacher model (LLM) reviews the student model's explanations, corrects inaccuracies, and refines the reasoning process. This direct intervention helps transfer not only knowledge but also critical reasoning skills to the student model, enabling it to perform robustly on its own. These methods collectively aim to optimize the knowledge distillation process by prioritizing the quality of reasoning and explanation. This focus helps overcome common issues in traditional distillation, such as inefficient learning and incomplete knowledge transfer, by ensuring that student models learn to reason accurately and efficiently, thereby enhancing their independent problem-solving capabilities.



Customized LLM-Annotated Explanation Revision

Prompt: For question <q>. Let's think step by step. Can we start with this step: <s_1>? If yes, give me your step. If no, give me the rest steps and the final answer.
Response: Yes, we can start with that <$\hat{s}_1$>
Prompt: Can we continue with this step <s_2>?
Response: Yes, the second step is <$\hat{s}_2$>.
Prompt: Can we continue with this step <s_3>?
Response: No, we should proceed as <$\hat{s}_{\geq 3}$>, the final answer is <$\hat{a}$>.

Results

In evaluating the ELAD framework, significant performance enhancements were noted across various reasoning tasks when compared with traditional methods. The Explanation-Guided Sample Selection (EGSS) method outperformed baseline sample selection techniques, showing notable improvements such as a 2.41% increase in accuracy for the GSM8K dataset and a 3.27% boost for the ANLI dataset. Additionally, the Customized LLM-Annotated Explanation Revision (CLEAR) method demonstrated superior effectiveness over conventional CoT prompting, with accuracy improvements of up to 2.71% in GSM8K and 5.31% in StrategyQA, confirming the framework's efficacy in refining reasoning skills and selecting highly informative samples as annotation budgets increase.

Method	Annotating	*Arithmetic*		*NLI*		*Commonsense*	
		GSM8K	AQuA	ANLI	e-SNLI	CommonSenseQA	StrategyQA
Teacher: `GPT-3.5-turbo`							
Zero-shot-CoT	–	73.45	54.96	68.02	47.67	68.94	69.78
Student: `LLaMA-2-7B`							
Zero-shot-CoT	–	10.04	21.07	33.94	28.98	41.28	44.71
Fine-Tuned Student							
Random	CoT Prompting	28.42	26.86	54.22	48.60	45.66	48.76
	CLAER	30.31	27.05	57.12	48.56	48.54	50.89
Maximum Entropy	CoT Prompting	27.58	27.67	52.56	47.98	46.35	49.03
	CLAER	29.04	27.42	53.75	51.76	48.86	51.05
Least Confidence	CoT Prompting	28.42	25.8	52.26	48.21	45.93	47.53
	CLAER	28.68	27.19	53.63	48.65	48.52	51.23
Disagreement	CoT Prompting	30.11	25.91	55.59	50.32	48.64	48.60
	CLAER	31.49	27.23	58.71	54.32	52.46	53.81
Self-Confidence	CoT Prompting	26.41	26.04	52.69	46.01	48.53	49.69
	CLAER	27.95	25.57	54.32	49.21	49.03	52.44
EGSS	CoT Prompting	30.01	26.91	55.87	51.16	49.64	50.32
	CLAER	32.72	28.43	58.02	54.44	53.53	55.63

Conclusions

This paper introduced the Explanation-Guided LLMs Active Distillation (ELAD) framework to address the challenges of deploying LLMs due to the high memory and computational demands. Our proposed framework achieves LLMs active distillation with explanation-guided sample selection and a customized LLM-annotated explanation revision. Extensive experiments on various reasoning datasets demonstrate the effectiveness of our approach in enhancing the distillation efficiency.

Contact

Yifei Zhang: yifei.zhang2@emory.edu
Bo Pan: bo.pan@emory.edu
Chen Ling: chen.ling@emory.edu
Yuntong Hu: yuntong.hu@emory.edu
Dr. Liang Zhao: liang.zhao@emory.edu

References

1. Guangji Bai, Zheng Chai, Chen Ling, Shiyu Wang, Jiaying Lu, Nan Zhang, Tingwei Shi, Ziyang Yu, Mengdan Zhu, Yifei Zhang, et al. 2024. Beyond efficiency: A systematic survey of resource-efficient large language models. arXiv preprint arXiv:2401.00625.
2. Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168.
3. Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.
4. Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. arXiv preprint arXiv:2203.15556.
5. Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2019. Tinybert: Distilling bert for natural language understanding. arXiv preprint arXiv:1909.10351.
6. Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. 2023. Symbolic chain-of-thought distillation: Small models can also" think" step-by-step. arXiv preprint arXiv:2306.14050.
7. Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2022. Teaching small language models to reason. arXiv preprint arXiv:2212.08410.
8. Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.