L³ Ensembles: Lifelong Learning Approach for Ensemble of Foundational Language Models*

Aidin Shiri[1], Kaushik Roy[2], Amit Sheth[2], Manas Gaur[1]

[1]University of Maryland, Baltimore County (UMBC), MD, USA; [2] AI Institute, University of South Carolina, SC, USA

{aidin.shiri, manas}@umbc.edu, kaushikr@email.sc.edu, amit@sc.edu

1 ABSTRACT

[1] Fine-tuning pre-trained foundational language models (FLM) for specific tasks is often impractical, especially for resource-constrained devices. This necessitates the development of a Lifelong Learning (L³) framework that continuously adapts to a stream of Natural Language Processing (NLP) tasks efficiently. We propose an approach that focuses on extracting meaningful representations from unseen data, constructing a structured knowledge base, and improving task performance incrementally. We conducted experiments on various NLP tasks to validate its effectiveness, including benchmarks like GLUE and SuperGLUE. We measured good performance across the accuracy, training efficiency, and knowledge transfer metrics. Initial experimental results show that the proposed L³ ensemble method increases the model accuracy 4%~36% compared to the fine-tuned FLM. Furthermore, L³ model outperforms naive fine-tuning approaches while maintaining competitive or superior performance (up to 15.4% increase in accuracy) compared to the state-of-the-art language model (T5) for the given task, STS benchmark.

2 INTRODUCTION

When training models in Artificial Intelligence (AI) and Machine Learning (ML), the capacity for models to continually learn and adapt to new tasks and data distributions is a critical challenge. Typically, AI and ML models are meticulously trained on specific datasets to acquire domain knowledge, expecting that this knowledge can then be applied to perform well on unseen but related data. However, real-world applications often demand more flexibility. These applications require models to learn new tasks efficiently and retain the knowledge of previously learned tasks, thus avoiding what is known as catastrophic forgetting (CF). CF occurs when adapting a model to a new task leads to a significant loss of knowledge in previously learned tasks.

In recent years, the research focus has shifted towards addressing this issue through the paradigm of **L**ife**L**ong **L**earning, or L³ [6]. Existing L³ approaches have explored various strategies, including regularization, model architecture, and data-based techniques. For instance, some methods employ regularization techniques to consolidate weights associated with previous tasks when learning new ones [3]. Others isolate specific model parameters for different tasks or incorporate replay-based approaches using old task data to guide new task learning [1][9]. Within the realm of Natural Language Processing (NLP), especially in resource-constrained scenarios like edge devices, there is a greater demand for straightforward and efficient Foundational Language Models (FLMs). This demand arises when comparing them to their more intricate counterparts with over a billion parameters. The practical applications of FLMs in

Figure 1: Performance (%) of the Fine-Tuned $BERT_{base}$ on Continual Sequence of Similar GLUE Benchmark Tasks (C: CoLA, S: SST2, QQ: QQP, W: WNLI, QN: QNLI). Notation "$C \rightarrow S$" denotes training on CoLA and testing on SST2. Notation "$QQ \rightarrow W \rightarrow QN$" indicates that the model underwent training on QQP, followed by WNLI, and was subsequently tested on QNLI.

Figure 2: A framework for Lifelong Learning using Ensemble of FLMs. Blue tasks are the parts that have been implemented, and green parts are works in progress.

edge devices encompass tasks like question-answering, engaging in conversations, and extracting information from visual content (e.g., named entities)[6][2].

This paper argues that rather than elevating the complexity of FLMs, more favorable outcomes can be attained through the fusion of multiple simpler FLMs enriched with infused knowledge.

This paper proposes an alternative approach to L³ that leverages ensemble configurations of pre-trained models and external knowledge augmentation to combat CF without requiring compute-intensive techniques. A compelling demonstration of the concept of CF in NLP can be observed in Figure 1. It vividly illustrates how, when we train $BERT_{base}$ on a specific task and then assess its performance on a semantically related dataset, we witness a notable decline in its performance. Furthermore, the same phenomenon occurs when $BERT_{base}$ is trained on two similar datasets; it still experiences a decrease in performance when tested on a third dataset, even if that dataset involves a related task. This striking evidence

Task	Step	FineTuned Model	Evaluation Task	Accuracy (A)	Comparison Baseline (CB)	Knowledge Transfer ($A - CB$)
GLUE TASKS: QQP and MRPC	1	$BERT_{base \to QQP}$	QQP	91.8%	$BERT_{base}$: 63%	28.8% (+)
	2	$BERT_{QQP}$	MRPC	71.8%	$BERT_{base}$: 31.6%	40.2% (+)
	3	$BERT_{QQP \to MRPC}$	MRPC	86.2%	$BERT_{QQP}$: 71.8%	14.4% (+)
	4	$BERT_{MRPC}$	QQP	84.9%	$BERT_{QQP}$: 91.8%	**CF: 7%** (-)
SuperGLUE TASKS: BoolQ and RTE	1	$BERT_{base \to RTE}$	RTE	72.5	$BERT_{base}$: 47.2%	25.3% (+)
	2	$BERT_{RTE}$	BoolQ	60%	$BERT_{base}$: 37.8%	22.2% (+)
	3	$BERT_{RTE \to BoolQ}$	BoolQ	75.2%	$BERT_{RTE}$: 60%	15.2% (+)
	4	$BERT_{BoolQ}$:	RTE	43.6%	$BERT_{RTE}$: 72.5%	**CF: 28.9%** (-)

Table 1: The Knowledge transfer of single FLM "$BERT_{base}$" on the four sample GLUE and SuperGLUE tasks. Notation "$BERT_{base \to QQP}$" denotes that the base model is fine-tuned on QQP task.

FLM	Size	MSE	Ensemble	Size	Naïve Ensemble	Weighted Ensemble	LLM Ensemble	KI Ensemble
BERT	110M	0.39	BERT & DistilBERT	176M	0.382	0.382	0.393	**0.374**
DistilBERT	66M	0.45	BERT & ELECTRA	220M	0.304	0.302	**0.285**	0.288
ELECTRA	110M	0.34	BERT & RoBERTa	235M	0.294	0.290	**0.275**	0.293
RoBERTa	125M	0.32	BERT & DistilBERT & ELECTRA	286M	0.301	0.316	0.312	**0.295**
T5	11B	0.31	BERT& RoBERTa & ELECTRA	345M	0.286	0.282	**0. 262**	0.264

Table 2: Effect of Ensembling and Knowledge Infusion on the MSE loss of the STS benchmark dataset. We see that ensembling improves performance even with models of modest size - this is especially noteworthy in the last row with the T5 model.

underscores the significance of addressing CF challenges in NLP for usability in edge devices.

Traditional ensembling often involves simple or weighted aggregation methods, which can lead to suboptimal performance when solving tasks. Naive ensemble and weighted ensemble are previously proposed [4]. We introduce the Large Language Model (LLM) Ensemble and Knowledge Infused (KI) ensemble as two methods to prevent CF in FLM training or fine-tuning. In the *LLM Ensemble*, we harness the power of frozen embeddings from LLMs to enhance vector representations through meticulous modulation. Specifically, we incorporate embeddings derived from "Langchain text-embedding-ada-002 LLM" into our LLM Ensemble approach, a pivotal step in fortifying the model's vectorized representations. For the *KI Ensemble*, as showcased in Figure 2, we capitalize on the vectorized information from Wikipedia Knowledge Graph.

3 EXPERIMENTS AND DISCUSSION

Tasks and Datasets: We employ multiple datasets from two established benchmarks in our experiments. One of these benchmarks is the General Language Understanding Evaluation (GLUE), encompassing a variety of natural language understanding tasks [8]. The second is an enhancement over GLUE called SuperGLUE, which includes a relatively more demanding and varied assortment of tasks [7]. ***Individual Model Baselines.*** As baselines, we utilize various FLMs, including BERT, RoBERTa, DistilBERT, and ELECTRA. We first assess individual FLM performance on various GLUE tasks, as illustrated in Figure 1 and Table 1.

Knowledge Transfer and Catastrophic Forgetting of Individual Models. We measure the individual FLM model performance on a sequence of tasks (using accuracy). Our findings indicate that while fine-tuning boosts performance on individual tasks (illustrated by base −> task in **Table 1.**), training a model fine-tuned on one task for another task causes CF of the old task (between 7% and 28.9%). Thus, an ensemble approach has the potential to mitigate CF while enhancing knowledge transfer.

Ensemble Methods for Improving Performance in Resource-Constrained Systems. Next, we experiment with another dataset, STS, which is most representative of the tasks across the benchmark datasets. This time, we report the results of individual FLM performance vs. different ensemble configurations. We find that the ensemble performs better than the individual models, and crucially, all of the individual models are at most 500M in size - showing that ensemble methods are a good choice for resource-constrained domains (see Table 2) [2].

4 CONCLUSION AND FUTURE WORK

Fine-tuning FLMs improved task-specific performance in GLUE tasks, but transferring a fine-tuned model to another task led to a significant performance drop due to CF. Our work highlighted the need for a L^3 ensemble approach to mitigate this issue, demonstrating the superior performance of ensembles over individual models. The findings emphasize the potential of ensemble methods to enhance knowledge transfer and address CF in settings where model size and efficiency are crucial for success, such as resource-constrained settings. This research will extend KI Ensemble using a reinforcement learning-based approach (see Figure 2, green boxes) across various knowledge-intensive NLP tasks [5].

REFERENCES

[1] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. 2019. Continual learning: A comparative study on how to defy forgetting in classification tasks. *arXiv preprint arXiv:1909.08383* (2019).

[2] Kalpa Gunaratna, Vijay Srinivasan, Sandeep Nama, and Hongxia Jin. 2021. Using neighborhood context to improve information extraction from visual documents captured on mobile phones. In *ACM CIKM*.

[3] Sang-Woo Lee, Jin-Hwa Kim, Jaehyun Jun, Jung-Woo Ha, and Byoung-Tak Zhang. 2017. Overcoming catastrophic forgetting by incremental moment matching. *NIPS* (2017).

[4] Michael S Matena and Colin A Raffel. 2022. Merging models with fisher-weighted averaging. *NIPS* (2022).

[5] Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. 2021. KILT: a Benchmark for Knowledge Intensive Language Tasks. In *NAACL*.

[6] Fan-Keng Sun, Cheng-Hao Ho, and Hung-Yi Lee. 2019. Lamol: Language modeling for lifelong language learning. *arXiv preprint arXiv:1909.03329* (2019).

[7] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *NIPS* (2019).

[8] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461* (2018).

[9] Zirui Wang, Sanket Vaibhav Mehta, Barnabás Póczos, and Jaime Carbonell. 2020. Efficient meta lifelong-learning with limited memory. *arXiv preprint arXiv:2010.02500* (2020).