EXPLORING THE DLTH IN FINETUNING THROUGH SPECIALISED SPARSIFICATION

Sampreeth R S
Department of Computer Science
Indian Institute of Technology
Kharagpur, India 577204.
sampreeth2003@kgpian.iitkgp.ac.in

Arindam Biswas
Research Scientist
Polynom
Paris, France.
arindam.biswas@polynom.io

Pabitra Mitra
Department of Computer Science
Indian Institute of Technology
Kharagpur, India 577204.
pabitra@cse.iitkgp.ac.in

Biswajit Basu
School of Engineering
Trinity College Dublin
Dublin 2, Ireland
basub@tcd.ie

ABSTRACT

Adapting foundation models to new tasks often involves modifying all model weights, leading to destructive interference such as catastrophic forgetting and degraded multi-task performance. Sparse adaptation methods like Lottery Ticket Adaptation (LoTA) mitigate these issues by optimizing only sparse subnetworks, achieving better results and enabling model merging across dissimilar tasks. Concurrently, the Dual Lottery Ticket Hypothesis (DLTH) states that randomly selected subnetworks can be transformed to a trainable condition that matches the performance of winning tickets. In this work, our goal is to explore the DLTH in sparse transformer finetuning tasks. We introduce a novel approach that employs expander graph masks to obtain an initial sparse subnetwork instead of random selection. In the first stage by maintaining a high spectral gap through expander masks, we transform randomly selected subnetworks into trainable ones. This method not only improves accuracy over random pruning but also uses the same mask across all layers, simplifying the adaptation process. This approach demonstrates expander-based initial pruning enhances sparse adaptations in foundation models, with the potential of addressing multi-task learning challenges without destructive interference.

1 INTRODUCTION

Overparameterized neural networks have achieved remarkable success across various machine learning tasks. However, their significant computational and storage demands present considerable resource constraints. Pruning techniques have emerged to address this issue by eliminating less important weights, resulting in sparse networks that are more efficient but often experience performance degradation. The Lottery Ticket Hypothesis (LTH) Frankle & Carbin (2018) provides a compelling perspective by suggesting that within a randomly initialized dense network, there exists a sparse subnetwork–a *"winning ticket"*–that can be trained to match the performance of the original network. While LTH focuses on finding these winning tickets through specific pruning methods, it overlooks the potential of random subnetworks within the dense network. Addressing this gap, the Dual Lottery Ticket Hypothesis (DLTH) proposes that randomly selected subnetworks can be transformed into trainable ones, effectively turning *"random tickets"* into winning tickets (Xu & Zhang, 2024).

In another direction, adapting language models to new tasks often involves updating all model weights, which can lead to issues like catastrophic forgetting and degraded performance in multi-task learning scenarios (Hu et al., 2021; Han et al., 2024). Lottery Ticket Adaptation (LoTA) Panda et al. (2024) has been proposed to mitigate these problems by optimizing only a sparse subnetwork

within models like RoBERTa Liu (2019), thereby preserving performance across multiple tasks and enabling model merging over dissimilar tasks. However, LoTA initially relies on randomly pruned subnetworks, which may not capture the most effective sparse structures for training.

This gives rise to the question of how we can exploit the LTH and DLTH to obtain sparse masks which can in turn be trained on new tasks to mitigate the effects of catastrophic forgetting and degraded performance. In this work, we establish the DLTH in the context of finetuning by introducing a new approach that utilizes *expander masks* for pruning instead of random selection. By carefully choosing our masks within language models like RoBERTa and Llama 3 we enhance adaptation accuracy and efficiency. Notably, the same mask is used across all layers, simplifying the adaptation process and reducing complexity.

Our conducted experiments demonstrate improvements over traditional random pruning and then fientuning methods. The use of expander masks leads to better accuracy and more efficient sparse adaptations, effectively addressing multi-task learning challenges without the detrimental effects of destructive interference.

Our contributions are summarized as follows:

1. **Establishing DLTH for finetuning with expander masks on RoBERTa and Llama Models**: We validate the Dual Lottery Ticket Hypothesis within the framework of Lottery Ticket Adaptation by employing expander masks for pruning in RoBERTa and Llama models, enhancing the trainability of randomly selected subnetworks.
2. **Improved Accuracy and Efficiency**: By maintaining a high spectral gap through expander masks, our method outperforms random pruning techniques in accuracy while using a consistent mask across all network layers.
3. **Advancements in Sparse Adaptation for Language Models**: This approach enhances sparse network training and adaptation methods, with the goal of mitigating issues like catastrophic forgetting and destructive interference in multi-task learning scenarios.

2 Background and Methodology

We briefly explain some of the background details and terms we have used in this work. Further details can be found in the appendix A and in the references mentioned therein.

1. **Expander** An expander is a highly connected and yet sparse graph structure which allows efficient flow of information in between the nodes.
2. **Lottery tickets** Lottery tickets refer to subnetworks within dense neural networks that can be pruned yet still achieve high performance when trained independently.
3. **Dual lottery tickets** Dual lottery tickets extend the original Lottery Ticket Hypothesis by proposing that randomly selected subnetworks from a dense, randomly initialized network can be transformed into trainable structures.

Now we describe the sparse network masks used in our approach. The method of sequential finetuning using the sparse subnetworks obtained by applying the mask is described next.

2.1 Random Masks

Random masks are applied to model parameters with a specified sparsity, focusing on the query, key, and value matrices of each layer of the transformer models. For each task the model needs to be finetuned on, a distinct adapter is trained. These adapters use random sparse networks to tailor the model's parameters to the specific task. The adapter weights pertaining to the mask are zeroed out using RST and only the non-masked weights are updated in order to fine-tune the model for the task at hand. Hence only a small portion of the model weights are updated for the task while maintaining performance close to or better than full fine tuning. By training a separate adapter for each task, we maintain task-specific adjustments while preserving the base model's shared knowledge. The fine-tuning process is conducted by only modifying the weights within the mask. This enables us to obtain a dual lottery ticket, where a sparse subnetwork, identified by the random masks, can achieve

performance comparable to that of the dense model. The combination of sparse network masking and task-specific adapter training ensures that each task is optimally finetuned within the sparsity constraints.

2.2 EXPANDER MASKS

We move from random masking to using structured masks derived from expander graphs. Expander graphs are known for their excellent sparsity and connectivity properties, which make them highly suitable for constructing efficient sparse networks. To generate the masks, we treat the bi-adjacency matrix of the stacked bipartite expander graph as the mask itself. This bi-adjacency matrix inherently captures the sparse yet highly connected nature of the expander graph. We apply the same adjacency matrix-derived mask uniformly across all layers and across the query, key, and value matrices. This approach allows us to investigate whether structured sparsity, imposed by expander graph connectivity, leads to more efficient fine-tuning without compromising model performance. By using a consistent mask across the model, we also explore how uniform sparse subnetworks affect the overall generalization capabilities of the model across different tasks. Recently, there has been a flurry works on expander masks and lottery tickets.

2.3 SEQUENTIAL TRAINING

In our sequential training approach, we extend the use of expander graphs to handle multiple tasks in a sequential manner. To achieve this, we choose two distinct expander graphs whose edges have very small intersection ($< 10\%$). This ensures that the masks formed from these graphs do not overlap much, maintaining separate sparse subnetworks for each task.

The training begins with the first task, where the mask for this task is derived from the first expander graph. After completing the training on the first task, the weights corresponding to the masked connections are frozen, preserving the learned parameters for that task.

For the second task, we introduce the mask formed from the second expander graph, ensuring that it does not interfere with the previously learned weights. This sequential training method allows the model to adapt to multiple tasks while preserving knowledge from earlier tasks, thanks to the mutually exclusive nature of the expander graph-derived masks.

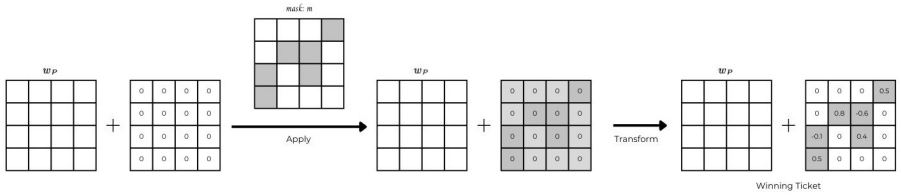

Figure 1: (1) Mask calibration using Random masks or Expander graphs. (2) Subnetwork extraction using RST. (3) Winning Lottery Ticket (wp represents the base weights of the model)

3 EXPERIMENTS

3.1 GLUE TASKS

In order to evaluate the effectiveness of our method, we conducted experiments using the RoBERTa model on a subset of tasks from the GLUE (General Language Understanding Evaluation) benchmark. Our experiments were carried out on both the RoBERTa base and RoBERTa large models to examine the impact of model size on performance. For all experiments, we used a batch size of 16 and a learning rate of 1×10^{-4}. The tasks included in our evaluation are CoLA, RTE, MRPC, STS-B, SST-2, and QNLI. Due to time constraints, we omitted SST-2 and QNLI from the RoBERTa large model, as well as the MNLI and QQP tasks from our evaluation.

3.2 GENERATIVE TASKS

We evaluate our methodology by training meta-llama/Meta-Llama-3-8B model on five fronts: instruction following, mathematics, programming, summarisation, and reasoning. We will now succinctly examine each capacity, the datasets utilised for fine-tuning and evaluating the proposed methodologies, and the rationale for their selection.

Instruction following: For this purpose, we train models to data from UltraFeedback (Cui et al., 2024), which encompasses a variety of data points addressing truthfulness, honesty, and helpfulness, with instruction adherence. We assess the instruction-following capability using the length-controlled AlpacaEval2 Win Rate (Dubois et al., 2025), which we denote as "winrate". A high success rate indicates that GPT-4 favours the replies generated by our model on a selection of typical prompts compared to its own responses. Winrate is the measure most strongly associated with human rating preference.

Reasoning: We train our model on 8 reasoning datasets: Boolq (Clark et al., 2019), PIQA (Bisk et al., 2019), SocialQA (Sap et al., 2019), Hellaswag (Zellers et al., 2019), Winogrande (ai2, 2019), Arc-Easy (Clark et al., 2018), Arc-Challenge (Clark et al., 2018), OpenBookQA (Mihaylov et al., 2018) and report the exact match accuracy.

Math: We train our model on the gsm8k (Cobbe et al., 2021) dataset and report the accuracy of the model.

Code generation: We use the SQL-create-context (b mc2, 2023) dataset that contains instruction prompts for the model to write SQL queries and report the ROUGE-F1 score on the test set.

Summarization: We utilise Samsum (Gliwa et al., 2019) dataset and present the ROUGE-1 F1 score for the test set.

3.3 RESULTS

We first compare the performances of random masks and expander masks for the RoBERTa base (Table 1), and RoBERTa large (Table 2) models on the GLUE tasks. The sparsity, denoting the fraction of masked parameters, is set to 99%. It can be observed the expander mask outperforms the random mask for all the tasks.

Table 1: Results on GLUE Tasks on RoBERTa Base with 99% Sparsity

Task	CoLA	RTE	MRPC	STS-B	SST-2	QNLI
Random Mask	0.244	0.559	0.828	0.876	0.926	0.893
Expander Mask	0.566	0.720	0.833	0.896	0.928	0.916

Table 2: Performance on GLUE Tasks on RoBERTa large with 99% Sparsity

Task	CoLA	RTE	MRPC	STS-B
Random Mask	0.655	0.815	0.889	0.911
Expander Mask	0.677	0.837	0.892	0.914

Next, we study viability of sequential training on a pair of tasks while using the expander masks. The sequential tasks were performed with a sparsity of 90% on the RoBERTa base model achieved by applying the expander masks. Table 3 show the performance metrics for individual tasks in a task pair after the model was finetuned sequentially for the task pair. We observe a slight drop in the performance of the Task 1 in each case, but not a catastrophic degradation.

Table 3: Performance on sequential training on RoBERTa Base with 99% sparsity obtained using expander masks. (Task-1 trained first followed by Task-2. Evaluation metrics computed for each task after training on both tasks.)

Tasks	Task-1 metric	Task-2 metric
MRPC-CoLA	0.686	0.570
RTE-MRPC	0.498	0.867
CoLA-RTE	0.109	0.776
CoLA-MRPC	0.160	0.877
MRPC-RTE	0.344	0.758
RTE-CoLA	0.462	0.572

Following this, we apply our methodology to the meta-llama / meta-lama-3-8B model to train the model on the tasks mentioned earlier, and the results are shown in table 4. The results for LoTA and our method use a mask with 10% sparsity and a learning rate of 1e-6, whereas the hyperparameters used for LoRA have been taken from Panda et al. (2024)

Table 4: Performance of meta-llama/Meta-Llama-3-8B model on various tasks using expander masks with 10% sparsity, best results have been mentioned in bold.

Task	FFT	LoRA	LoTA	Our Method
gsm8k	63.4	62.3	63.2	**66.4**
Reasoning	84.8	84.1	84.4	**98.53**
SQL	**99.4**	98.7	99.0	98.9
Summarisation	53.6	52.3	52.3	**54.8**
Instruction following	17.61	14.2	**18.0**	14.9

4 Conclusion, limitations and future work

In this work, we explored the Dual Lottery Ticket Hypothesis (DLTH) in the context of finetuning, with specialised sparsification as an effective alternative to random pruning. Expander masking is found to improve sparse adaptation methods in foundation models like RoBERTa and Llama. Our results demonstrate that expander sparsification, improve both accuracy and parameter efficiency over random masking methods, especially in scenarios involving high sparsity. Additionally, we showed that it is possible to maintain task-specific learning while mitigating issues like catastrophic forgetting and destructive interference in multi-task learning.

There are several avenues for future work that we aim to explore. One of the primary directions is extending our methodology to vision models, other LLMs and MLLMs. The current findings suggest that expander-based pruning could generalize well to these larger architectures, potentially providing a solution to the growing computational complexity and efficiency demands in adapting LLMs to multiple tasks. We also plan to further investigate the impact of structured sparsity on more complex multi-task learning scenarios, as well as improve the sequential training process by refining mask calibration techniques.

References

Winogrande: An adversarial winograd schema challenge at scale. 2019.

Noga Alon. Eigenvalues and expanders. *Combinatorica*, 6(2):83–96, Jun 1986. ISSN 1439-6912. doi: 10.1007/BF02579166. URL https://doi.org/10.1007/BF02579166.

b mc2. sql-create-context dataset, 2023. URL https://huggingface.co/datasets/b-mc2/sql-create-context. This dataset was created by modifying data from the following sources: Zhong et al. (2017); Yu et al. (2018).

Yue Bai, Huan Wang, ZHIQIANG TAO, Kunpeng Li, and Yun Fu. Dual lottery ticket hypothesis. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=fOsN52jn25l.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language, 2019. URL https://arxiv.org/abs/1911.11641.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions, 2019. URL https://arxiv.org/abs/1905.10044.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with scaled ai feedback, 2024. URL https://arxiv.org/abs/2310.01377.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators, 2025. URL https://arxiv.org/abs/2404.04475.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2018.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pp. 70–79, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5409. URL https://www.aclweb.org/anthology/D19-5409.

Zeyu Han, Chao Gao, Jinyang Liu, Sai Qian Zhang, et al. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*, 2024.

Shlomo Hoory, Nathan Linial, and Avi Wigderson. Expander graphs and their applications. *Bulletin of the American Mathematical Society*, 43(4):439–561, 2006.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021.

Yinhan Liu. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*, 2018.

Ashwinee Panda, Berivan Isik, Xiangyu Qi, Sanmi Koyejo, Tsachy Weissman, and Prateek Mittal. Lottery ticket adaptation: Mitigating destructive interference in LLMs. In *2nd Workshop on Advancing Neural Network Training: Computational Efficiency, Scalability, and Resource Optimization (WANT@ICML 2024)*, 2024. URL https://openreview.net/forum?id=qD2eFNvtw4.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialiqa: Commonsense reasoning about social interactions, 2019. URL https://arxiv.org/abs/1904.09728.

Jing Xu and Jingzhao Zhang. Random masking finds winning tickets for parameter efficient fine-tuning. In *International Conference on Machine Learning*, 2024.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, et al. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. *arXiv preprint arXiv:1809.08887*, 2018.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

Victor Zhong, Caiming Xiong, and Richard Socher. Seq2sql: Generating structured queries from natural language using reinforcement learning. *CoRR*, abs/1709.00103, 2017.

A Appendix

Expanders An expander is a highly connected and yet sparse graph structure which allows efficient flow of information in between the nodes. Intuitively, it is a finite, undirected multigraph in which every subset of the vertices that is not "too large" has a "large" boundary. This is mathematically quantified via the Cheeger constants. A graph $\mathbb{G} = (V, E)$ is an ϵ-vertex expander if for every non-empty subset $X \subset V$ with $|X| \leq \frac{|V|}{2}$, we have $\frac{|\delta(X)|}{|X|} \geq \epsilon$, where $\delta(X)$ denotes the outer vertex boundary of X i.e., the set of vertices in Γ which are connected to a vertex in X but do not lie in X. As X runs over all subsets of V, the infimum of $\frac{|\delta(X)|}{|X|}$ satisfying the conditions above is known as the vertex Cheeger constant and is denoted by $\mathfrak{h}(\Gamma)$. Thus, a large vertex Cheeger constant implies that it is difficult to disconnect the graph by small cuts. More details on expanders can be found in Alon (1986); Hoory et al. (2006) etc. In this work we shall mainly be concerned with bipartite expanders i.e., sparse bipartite graphs which are expanders.

Lottery tickets Lottery tickets refer to subnetworks within dense neural networks that can be pruned yet still achieve high performance when trained independently. This concept stems from the Lottery Ticket Hypothesis (LTH), which posits that a dense neural network contains sparse subnetworks ("winning tickets") that can perform as well as, or even better than, the original model. Finding lottery tickets involves training the dense network, pruning parameters based on their magnitude, and resetting the remaining weights to their original values for retraining. This iterative train-prune-retrain process helps identify the winning ticket with minimal complexity and high computational efficiency. See Frankle & Carbin (2018).

Dual lottery tickets Dual lottery tickets extend the original Lottery Ticket Hypothesis by proposing that randomly selected subnetworks from a dense, randomly initialized network can be transformed into trainable structures. This approach, called the Dual Lottery Ticket Hypothesis (DLTH), shifts the focus from identifying pre-existing winning tickets to transforming random subnetworks into high-performing ones. This transformation is achieved using regularization techniques that refine the selected subnetwork for improved performance. DLTH is crucial because it generalizes the sparse training process, offering flexibility and eliminating the need for pretraining dense networks, thus reducing computational costs while maintaining strong performance. See Bai et al. (2022).

Random Sparse Network Transformation (RST) RST is a training strategy aimed at transforming randomly selected sparse subnetworks into trainable structures by leveraging information from the rest of the network. The process involves applying a regularization term to extract useful information from the masked weights (the non-trainable part of the network) and transfer it to the unmasked weights (the trainable subnetwork). The regularization term gradually suppresses the magnitude of the masked weights while optimizing the overall loss function, ensuring that the unmasked weights receive the necessary information for training. As the training progresses, the importance of the masked weights diminishes, while the unmasked parameters become increasingly important. After sufficient training, the masked weights are removed, leaving a refined sparse network that is finetuned for final evaluation. In this way, the RST transforms a random subnetwork, or "random ticket", into a high-performing "winning ticket" by utilizing information from the entire network. See Bai et al. (2022) for details.