Comparative Study of LORA and (IA)^3 for Fine-Tuning DistilBERT on Extractive Question Answering

Tan Hwee Li Rachel

Github Link: Comparative-Study-of-LORA-and-IA-3-for-Fine-Tuning-DistilBERT-on-Extractive-Question-Answering

Abstract—The adaptation of large pretrained transformer models to downstream tasks is a central challenge in modern natural language processing, often hindered by the prohibitive computational cost of full fine-tuning. This work investigates the efficacy of Parameter-Efficient Fine-Tuning (PEFT) by comparing two distinct strategies, specifically Low-Rank Adaptation (LoRA) and Infused Adapter by Inhibiting and Amplifying Inner Activations ((IA)^3). A DistilBERT was fine-tuned for extractive question answering on the SQuAD v1 dataset using both methods. Although (IA)^3 demonstrated superior parameter efficiency, requiring substantially fewer trainable parameters, quantitative evaluation revealed that LoRA significantly outperformed (IA)^3. Qualitative error analysis further highlighted nuanced differences in model behavior. The findings conclude the importance of selecting an appropriate fine-tuning strategy for different applications.

I. Introduction

Large pretrained transformer models like BERT and its variants have become the foundation of modern NLP. However, their immense size makes full fine-tuning computationally expensive and resource-intensive. Parameter-Efficient Fine-Tuning (PEFT) methods have emerged as a powerful solution, enabling the adaptation of these models by training only a small fraction of their parameters.

This experiment aims to implement and compare two popular PEFT strategies to adapt a pretrained model for a specific downstream task. The chosen task is extractive Question Answering (QA), and the selected pretrained model is DistilBERT, a smaller, faster variant of BERT. By comparing Low-Rank Adaptation (LoRA) and (IA)^3, two methods that employ fundamentally different mechanisms for model adaptation, the goal is to analyse their performance differences, efficiency trade-offs, and qualitative behaviours to provide insights into their respective strengths and weaknesses.

II. Methodology

A. Dataset chosen

The dataset chosen for this experiment is the Standard Question Answering Dataset (SQuAD), specifically version 1.1. which is a standard benchmark for extractive QA.

SQuAD v1.1 was deliberately chosen over the more complex SQuAD v2.0 to align with the project's core objectives and computational constraints. In SQuAD v1.1, the answer to every question is guaranteed to be a segment of text within the provided context. In contrast, SQuAD v2.0 introduces questions that are unanswerable from the context, which adds the additional challenge of classifying answerability.

By focusing on version 1.1, this project could focus on the primary goal where it can directly compare the effectiveness of LoRA and (IA)^3 at the core task of information extraction, without the confounding variable of answerability detection. This was a pragmatic choice to ensure the experiments could be completed and analyzed effectively within a limited computational budget.

B. Pretrained Model

The uncased DistilBERT model was selected for this experimental task. The choice of DistilBERT over the larger BERT architecture was driven primarily by considerations of computational resource efficiency, facilitating a more rapid and streamlined computation. [2]

The 'uncased' version was utilized to standardize all textual input, converting both the dataset and input queries to lowercase prior to tokenization (e.g. what is a plant vs What is a plant?). This methodological decision aligns with best practices for general (QA) tasks where semantic meaning takes precedence over lexical casing, thereby establishing a robust and efficient baseline for the experiment.

C. Fine-Tuning Methods

The two fine-tuning methods chosen are both PEFT strategies, specifically Low-Rank Adaption (LoRA) and Infused Attention (IA)^3.

Low-Rank Adaptation (LoRA) is predicated on the hypothesis that the change in a model's weight matrix during fine-tuning exhibits a low intrinsic rank.[3] LoRA operates by freezing the original pretrained model weights and injecting a pair of trainable low-rank decomposition matrices into the self-attention mechanism of each transformer layer. These new matrices are optimized to adapt the model, and their outputs are subsequently added to the output of the corresponding original layer. For this project, LoRA was specifically applied to the query and value projection matrices within the self-attention blocks.

Infused Adapter by Inhibition and Amplifying Inner Activations (IA)^3 assumes that the weights frozen in pre-trained model is mostly sufficient. Hence, by learning to rescale inner activations across the transformer architecture, (IA)^3 incorporates small, trainable vectors that are multiplied element-wise with the outputs of predefined layers.[4] This mechanism effectively serves to either amplify or inhibit existing learned features. For the experiment, within the DistilBERT architecture, (IA)^3 was configured to rescale the activations in the key, query, value, and output layers of the attention block, as well as the intermediate and output layers of the feed-forward networks.

III. Experimental setup

A. Data Preprocessing

The raw text from the SQuAD training and validation sets was tokenized to align with the DistilBERT vocabulary. To handle contexts exceeding the model's maximum sequence length of 256 tokens, a sliding window approach was implemented. Long contexts were systematically split into shorter, overlapping chunks using a stride of 64 tokens. This process ensures complete coverage of the context while maintaining a manageable input size for the model. The tokenizer also generated offset mappings to trace tokens back to their original positions in the context, a critical step for accurately locating answer spans.

B. Hyperparameter Sweep

Prior to the final training phase, a hyperparameter sweep was conducted, focusing on the learning rate. This hyperparameter is critical as it fundamentally influences the model's convergence behavior. A learning rate that is too small can result in impractically slow convergence, whereas one that is too large may cause the training process to become unstable and diverge. To ensure computational tractability, the sweep was performed for a single epoch on a 10% subset of the training data. We evaluated learning rates from the set [$5e^{-5}$, $3e^{-5}$, $1e^{-5}$], selected to balance efficiency and performance. For each PEFT method, the learning rate that yielded the lowest validation loss was subsequently used for the final training run.

C. Final Training

Both LoRA and $(IA)^3$ models were trained for 3 epoch on the full training set using the learning rate found during the sweep.

Throughout this process, both training and validation loss were recorded at the end of each epoch. These metrics were subsequently plotted as learning curves to visualize the training dynamics. Furthermore, the total training time for each model was measured. This serves as a practical metric of computational efficiency, a primary evaluation criterion for PEFT methods.

D. Evaluation Metrics

Model performance was evaluated through both quantitative methods and qualitative methods. For quantitative evaluation, standard metrics such as Average Exact Match (EM) scores and Average F1 scores were calculated for each model.

The EM metric is a measure where each prediction receives a score of 100 if it perfectly matches the ground-truth answer after normalization, including lowercasing and removal of punctuation and articles and a score of 0 otherwise.

F1 score offers a more lenient evaluation by measuring the token-level overlap between the prediction and the ground truth. It is the harmonic mean of precision and recall, thereby accounting for partial matches. A higher F1 score indicates more token-level match and a lower indicates otherwise.

Beyond the standardized evaluation on the validation set, a comparative qualitative analysis was performed using custom-designed examples. This was done to probe the models' behaviors and gain deeper insights into the distinct strengths and weaknesses of each approach, supplementing the primary performance metrics.

IV. Results and evaluation

A. Hyperparameter Training Results

From Table 1, the hyperparameter learning rate of $5e^{-5}$ works the best for both LoRA and $(IA)^3$ models, achieving the lowest validation loss, specifically 3.556 and 4.802 respectively.

This finding is consistent with established best practices for fine-tuning Transformer-based models, where small learning rates in this range are widely recommended to ensure stable convergence and prevent the catastrophic forgetting of pre-trained knowledge. The empirical validation of this rate for both methods provided a solid foundation for the final training phase.

Method	Learning Rate	Training Loss	Validation Loss
LoRA	$5e^{-5}$	3.619	3.556
	$3e^{-5}$	3.873	3.761
	$1e^{-5}$	5.187	4.966
$(IA)^3$	$5e^{-5}$	4.965	4.802
	$3e^{-5}$	5.228	5.143
	$1e^{-5}$	5.443	5.423

Table 1: Table showing the training and validation loss for 1 epoch across different learning rate for both LoRA and $(IA)^3$ models

B. Final Training Results

Table 2 reveal a distinct trade-off between model performance and parameter efficiency for the LoRA and $(IA)^3$ methods. In terms of efficiency, the $(IA)^3$ model was exceptionally lightweight, requiring only 43,010 trainable parameters, which is approximately 6.9 times fewer than the 296,450 parameters needed by LoRA. Despite this significant reduction in parameter count, the training times were nearly identical, with LoRA taking 1.51 hours and $(IA)^3$ taking 1.54 hours. This indicates that primary advantage of $(IA)^3$ lies in model storage and memory efficiency rather than training speed.

Regarding performance, both models demonstrated successful learning, as shown by the consistent decrease in training loss and validation loss over three epochs, as shown in Figure 1 and 2. However, the LoRA model achieved a substantially better fit to the data, concluding with a final validation loss of 1.470, which was significantly lower than the 2.610 achieved by the $(IA)^3$ model. This suggests that while $(IA)^3$ is a more parameter-efficient approach, LoRA provides superior performance on this task as measured by convergence loss. Hence, it is expected for LoRA model to function better than $(IA)^3$ for the Question-Answering task.

Method	Trainable Parameters	Training Time /hours	Training Loss per Epoch	Validation Loss per Epoch
LoRA	296,450	1.51	2.010	1.775
			1.672	1.525
			1.570	1.470
$(IA)^3$	43,010	1.54	3.285	3.166
			2.841	2.710
			2.745	2.610

Table 2: Tabe showing the performance of LoRA and $(IA)^3$ models in terms of efficiency and loss

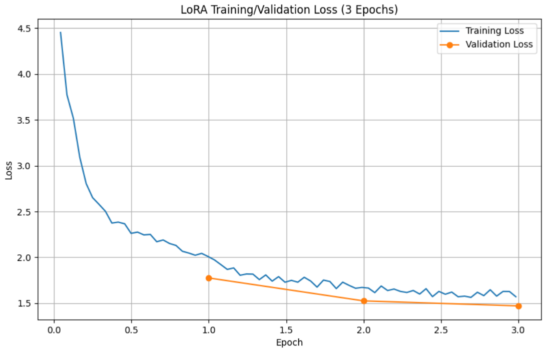

Figure 1: Training and Validation Loss for LoRA model across 3 epochs

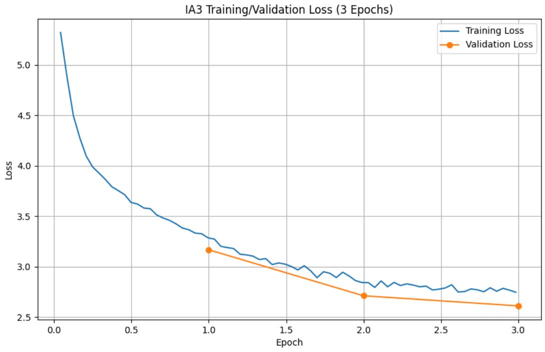

Figure 2: Training and Validation Loss for (IA)^3 model across 3 epochs

C. *Quantitative Results*

Table 3 shows the summary of evaluation metrics for both LoRA and (IA)^3 model. Looking at both the Average EM score and Average F1 score, LoRA has a higher score than (IA)^3, which suggests that it has a better performance in terms of retrieving the answer for the validation dataset in SQuAD. This superior performance aligns with the training dynamics observed previously, where LoRA consistently achieved a lower validation loss.

This outcome highlights a clear and important trade-off between parameter efficiency and task performance. While the (IA)^3 model is significantly more parameter-efficient, the additional trainable parameters afforded by LoRA appear to be crucial for capturing the more complex patterns required for the QA task. The performance gap indicates that this larger parametric budget translates directly into a tangible improvement in accuracy. Therefore, making LoRA a better model in this QA task.

Method	Average Exact Match (EM) Score	Average F1 Score
LoRA	62.3841	72.8802
(IA)^3	32.3652	41.8148

Table 3: Summary of evaluation metrics for both LoRA and (IA)^3 model

D. *Qualitative Results (Error Analysis)*

In terms of qualitative results, the custom examples have revealed nuanced behavioural difference. We will break down the findings into different cases targetting different aim.

i) Foundational Abilities

The models' foundational abilities in straightforward information extraction provided an initial baseline. The LoRA model demonstrated a robust capacity for semantic parsing, flawlessly identifying a defined term ("the frost line") within its contextual clause. This success suggests that its fine-tuning of attention weight matrices allows for a more nuanced understanding of syntactic and semantic relationships. In stark contrast, the (IA)^3 model exhibited a reliance on a shallower heuristic of token proximity. When tasked with the same question, it incorrectly extracted a list of entities ("mercury, venus, earth, and mars") that were merely adjacent to a keyword in the question ("rocky planets"), indicating a failure to grasp the sentence's definitional structure.

This weakness was further confirmed in a more complex and disambiguate task requiring anaphoric resolution. For instance, when resolving the task which involves tracing a pronoun back to its antecedent ("British Museum"), (IA)^3 failed to identify the correct entity type (a place) and instead extracted a date, revealing a less developed internal representation of semantic roles compared to LoRA, which succeeded.

ii) Temporal and Numerical Extraction with constraint

The analysis of temporal and numerical extraction tasks revealed a critical trade-off between specialized precision and generalized robustness, highlighting the distinct inductive biases imparted by the LoRA and (IA)^3 methods. This trade-off is most clearly exemplified by LoRA's paradoxical failure on a simple temporal query ("when did the solar system start?"). Despite its superior overall performance, the model returned "[No answer found]", suggesting a form of brittleness. This may be due to over-specialization on the complex linguistic patterns of the SQuAD dataset, creating a strong inductive bias that lowers the model's confidence for direct, simple queries that do not match its learned distribution.

In stark contrast, the (IA)^3 model demonstrated robustness on this simple query by correctly identifying the relevant context region. However, its success was undermined by a systematic deficiency in answer span delineation, as it returned an entire clause rather than the precise date. This pattern of imprecision recurred in other simple extractions, indicating that while (IA)^3's method of scaling internal activations is effective for identifying a semantically relevant passage, it is less adept at the fine-grained sequence boundary detection required for exactness.

This dynamic was inverted when the task complexity increased. A numerical extraction question requiring the model to parse multiple constraints and ignore several distractors showcased the strength of LoRA's approach. It successfully filtered the conditional logic of the query to isolate the correct value, with only a minor boundary error preventing a perfect exact match. This suggests that LoRA's low-rank adaptation of attention matrices provides a more sophisticated mechanism for handling complex, conditional queries. The (IA)^3 model, however, failed completely. Unable to process these layers of specificity, it was overwhelmed and defaulted to a no-answer prediction, revealing the limits of its simpler, more generalized approach when faced with highly constrained problems.

iii) Advanced Abilities

For multipart extraction, when faced with a question requiring the extraction of a multi-part answer, both models failed, as the task necessitates composing an answer from non-contiguous text spans which an operation outside the scope of extractive QA. The errors, however, were distinct, where the (IA)^3 model made a significant relational error, confusing the components of a merger with its outcome, while LoRA extracted an irrelevant sentence containing only one of the two required entities. An even clearer boundary was established by a causal inference question ("Why is Venus hotter than Mars?"). This task requires synthesizing information from multiple sentences to construct a new explanatory statement. Both models failed entirely, regressing to a primitive strategy of extracting tangentially related keywords ("melt lead," "carbon dioxide"). This complete failure in causal reasoning underscores a hard limit of the extractive paradigm itself, where models can retrieve what is stated but cannot logically infer what is implied.

V. Limitations

Firstly, the primary limitation of this experiment is the inherent constraint of the extractive question-answering paradigm itself. The models are architecturally bound to extracting a contiguous span of text from the provided context and are therefore incapable of tasks requiring abstractive summarization, synthesis of information from disparate passages, or true causal inference (which is also illustrated from the last Qualitative Analysis). Future work should extend this comparative analysis to generative or abstractive QA models, which are designed to overcome these fundamental limitations.

Second, the generalizability of our findings is constrained by the nature of the SQuAD dataset. Derived from Wikipedia, SQuAD features a formal, encyclopedic linguistic style and is known to contain lexical artifacts where high token overlap between question and context can lead models to learn superficial heuristics rather than deep comprehension. The brittleness observed in the LoRA model on simpler queries may be symptomatic of over-specialization on these dataset-specific patterns. Consequently, the relative performance of these PEFT methods may differ significantly on out-of-domain corpora, such as legal texts, clinical notes, or noisy social media data. Cross-domain evaluation remains a critical next step to assess the true robustness of these fine-tuning strategies.

Third, the scope of our experimental comparison, while focused, was not exhaustive. Due to computational constraints, this study did not include a traditional full fine-tuning baseline. Therefore, while we can speak to the relative merits of LoRA and (IA)^3, we cannot quantify their performance or efficiency trade-offs against the established, resource-intensive paradigm. Furthermore, the results are contingent on a limited search of the hyperparameter space for each PEFT method where only the learning rate is tuned. Hence, a more extensive search over specific hyperparameters such as the rank of LoRA could yield further and better results.

Finally, our evaluation relied on the standard metrics of Exact Match (EM) and F1-score. These metrics, while useful for benchmarking, are based on strict lexical overlap and fail to capture semantic equivalence. As our qualitative analysis revealed, a model can produce a semantically correct answer that is heavily penalized by these metrics. Future investigations would benefit from incorporating more advanced, model-based evaluation metrics that can better assess semantic fidelity and provide a more nuanced understanding of model performance.

VI. Key takeaways and conclusion

One of the key takeaways from this experiment is that LoRA demonstrates a clear and consistent performance advantage over (IA)^3 in terms of standard quantitative metrics for extractive question answering. However, our qualitative analysis reveals a more nuanced conclusion, which is the choice between these parameter-efficient fine-tuning (PEFT) methods involves a significant trade-off between specialized precision and generalized robustness. This study concludes that the optimal PEFT strategy is not absolute but is highly contingent upon the specific requirements of the downstream application.

The LoRA-tuned model manifested the profile of a specialist. Its architectural approach of adapting attention weight matrices enabled it to master complex linguistic patterns, resulting in superior performance on tasks requiring the parsing of multiple constraints, anaphoric resolution, and precise answer span delineation. This specialization, however, came at the trade off for brittleness. The model's unexpected failure on a simple, direct temporal query suggests an over-specialization on the complex patterns endemic to the SQuAD dataset, leading to a loss of robustness on out-of-distribution or simpler query structures. For applications demanding the highest possible precision within a well-defined domain, LoRA appears to be the superior method, provided that the target data closely mirrors the training distribution.

Conversely, the (IA)^3 model behaved as a generalist. Its method of scaling internal activations proved more robust for locating relevant information in response to simple, direct queries where LoRA failed. Yet, this robustness was consistently undermined by a systematic weakness in answer span detection, resulting in imprecise and verbose outputs that severely penalized its F1 scores. Furthermore, it was demonstrably overwhelmed by tasks requiring complex logical filtering, failing to return any answer. This suggests (IA)^3 may serve as a more resilient, albeit less precise, baseline for applications where simply identifying the correct region of context is sufficient.

In conclusion, this experiment successfully demonstrates that PEFT methods like LoRA and (IA)³ offer computationally efficient pathways to adapt large language models for specialized tasks. Our comparative analysis contributes to a deeper understanding of their distinct behavioural profiles, moving beyond aggregate performance metrics. While LoRA excels in precision and complex reasoning, its brittleness is a critical consideration. Ultimately, both methods are constrained by the fundamental limitations of the underlying extractive QA paradigm, failing at tasks that require abstractive reasoning or synthesis. This underscores that while the efficiency of fine-tuning has been largely democratized, the pursuit of true language understanding remains an ongoing challenge rooted in model architecture itself.

VII. Resources

[1] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016. [Online]. Available: https://arxiv.org/abs/1606.05250

[2] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," in *Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS 2019*, 2019. [Online]. Available: https://arxiv.org/abs/1910.01108

[3] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," in *International Conference on Learning Representations (ICLR)*, 2022. [Online]. Available: https://arxiv.org/abs/2106.09685

[4] J.-Y. He, A.-Z. Yen, and H.-H. Huang, "(IA)³: Infused Adapter by Inhibiting and Amplifying Inner Activations," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023. [Online]. Available: https://arxiv.org/abs/2205.05638