Summary of Changes and Revisions: for Paper "SC-LoRA: Balancing Efficient Fine-tuning and Knowledge Preservation via Subspace-Constrained LoRA"

Anonymous ACL submission

1 Overall Summary of Changes

We thank the reviewers and meta-reviewer for their constructive feedback. This revision addresses all major concerns through the following key changes:

1. **Core modification 1.** Corrected a written mistake when obtaining the covariance matrix in Algorithm 1, Section 3.2, where the tokens are stacked together.

2. **Core modification 2.** Modified related work part, compared more related methods.

3. **Core modification 3.** Clarifying the misunderstanding of risk of test data leakage in Section 4.1 and 4.2.

2 Point-by-Point Responses to Reviews

2.1 Meta-Reviewer

Suggested Revisions 1: The experiments are limited to a single model, llama2, and a few benchmarks, which makes it hard to know if the method works well on larger LLMs, languages, and other tasks like instruction following, reasoning. There's also a lack of comparisons with similar methods like Orthogonal Fine-Tuning, which makes it somehow harder to valid how well the proposed method performs.

Response: The number of benchmarks (3 fine-tuning tasks: math, summerization, math with poisoned data and 7 evaluation benchmarks: TriviaQA, NQ-open, WebQS, GSM8k, MATH, Samsum, and safety benchmark) is sufficient to demonstrate our method, on both fine-tuning performance and knowledge preservation. As for the model, we use llama2 and llama2-chat, which are general and widely-used by most researchers. Take CorDA (Yang et al., 2024) for example, they also apply their method on llama2 only.

The mentioned method Orthogonal Fine-Tuning is totally different from our LoRA-based method. And we focus on the initialization of LoRA, not the fine-tuning method itself. Main LoRA initialization methods are compared in our work.

Suggested Revisions 2: The paper claims to preserve a target subspace, but it doesn't show whether that subspace stays after training. Also, using only 256 samples to estimate covariance might weaken the results, an eval with more samples would make the findings more reliable.

Response: We do not restrict the fine-tuning on the subspace for it might weaken the expressiveness of the model and consequently the utility in the fine-tuning task. Also, previous work has shown the number of samples from 32 to 256 in estimating covariance are applicable.

Suggested Revisions 3: Some parts of the setup are unclear, especially how knowledge preservation is measured. Such info should be added in the revision to make it easier to read. Also, relying too much on subjective scoring from other models could affect the reliability of the results.

Response: World knowledge preservation is measured by exact match scores on common knowledge datasets including TriviaQA, NQ open, and WebQS, while the knowledge preservation of safety alignment is measured by attack success rate. For relying on subjective assessments, judging by LLM in safety tests is a well-developed technique, which has more accuracy than traditional Keywords searching based evaluation.

Summary Of Weaknesses 4: The paper would be easier to apply if it included a short "how-to" section with tuning tips, and fixing minor writing issues would help to make it more clear.

Response: For tuning the hyper-parameter, the experiments showed that setting β from 0.8 to 0.9 and number of samples in initialization by 256 shows best balanced performance in multiple tasks, while other hyper-parameters follows the same tuning techniques as vanilla LoRA. Paper writing have been polished in this revision.

Reviewer #NVEt

Summary Of Weaknesses 1: The authors claimed to pursue a trade-off between efficient fine-tuning and knowledge preservation. While no convergence analysis is presented in the paper.

Response: In Section 4.1 and 4.2, we follow the training hyper-parameters in previous works, where the fine-tuning tasks are guaranteed to converge.

Summary Of Weaknesses 2: The setting of preserving knowledge from pre-training data is problematic. As a task is specified, why do we use multi-task LoRA techniques to boost performance on both T^+ and T^- ? I think if we want to evaluate a fine-tuned LLM on its pre-training data, it would be better to fine-tune on a task A and test on another task B whose testing data is involved in the pre-training dataset.

Response: In many circumstances, boosting the ability on task T^- requires heavy training or huge amount of data, thus relying on multi-task LoRA techniques is computationally expensive, which motivated our research of light-weight and data-efficient approach.

Summary Of Weaknesses 3: For experiments shown in Sec. 4.1 and 4.3, the authors use data from the testing dataset to optimize the fine-tuned LLM. Does the experimental design risk test data leakage?

Response: We added explicit explanations to clarify that the data in initialization and training process are separate from the data in testing process.

Reviewer #mLAP

Summary Of Weaknesses 1: Knowledge preservation tests focus only on safety and general world knowledge, ignoring other aspects such as multimodal reasoning or contextual understanding. The authors claim that SC-LoRA outperforms competing methods but they rely on subjective assessments from the DeepSeek-V3 model, making it challenging to compare results with other studies.

Response: The experiment of fine-tuning on Samsum covers contextual understanding since the task requires the model to understand the conversation and give a summation. For relying on subjective assessments, judging by LLM in safety tests is a well-developed technique, which has more accuracy than traditional Keywords searching based evaluation.

Summary Of Weaknesses: The authors should verify their approach in broader setting, for different models and other (non-subjective) measures. Also a verification on different architecture (like ViT) could be interesting.

Response: This paper focuses on developing a new initialization method, and verified its generality on different tasks including summarization and math.

Reviewer #ArCT

Summary Of Weaknesses: In the paper, the authors use 256 samples to estimate the covariance matrices. A robustness analysis would strengthen the claims. And the samples are sequences of tokens. Are the hidden states of the tokens stacked together when computing the covariance matrices?

Response: Similar robustness analysis has already been done in previous work CorDA: Context-Oriented Decomposition Adaptation of Large Language Models for Task-Aware Parameter-Efficient Fine-tuning, Appendix B, showing number of samples from 32 to 256 are applicable. Unlike this work, we do not make use of inverse of covariance matrix, requiring less numerical precision. The hidden states of the tokens stacked together, and we have corrected the written mistake in Section 3.2, Algorithm 1 and Appendix B.

Summary Of Weaknesses: The initialization biases gradients toward updates that preserve the subspace S, but the constraints are not explicitly enforced during optimization. Is there any evidence of subspace preservation post-training? Comparison with methods like Orthogonal Fine-Tuning, which projects gradients onto the subspace S would be beneficial.

Response: We only constraint the weight in subspace at initialization, but not during the whole fine-tuning process, such restriction may limit the expressiveness of the fine-tuning model. Orthogonal Fine-Tuning adds a rotation matrix after each weight matrix, which differs from our LoRA-based method.

Summary Of Weaknesses: The experiments focus on safety/math. The performance on complex or multi-task fine-tuning (e.g., instruction-following) is not clear.

Response: We beleive that the number and variety of benchmarks (3 fine-tuning tasks: math, summerization, math with poisoned data and 7 evaluation benchmarks: TriviaQA, NQ-open, WebQS, GSM8k, MATH, Samsum, and safety benchmark) are sufficient to demonstrate the effectiveness of

our method, on both fine-tuning performance and knowledge preservation.

References

Yibo Yang, Xiaojie Li, Zhongzhu Zhou, Shuaiwen Leon Song, Jianlong Wu, Liqiang Nie, and Bernard Ghanem. 2024. Corda: Context-oriented decomposition adaptation of large language models for task-aware parameter-efficient fine-tuning. In *Advances in Neural Information Processing Systems*, volume 37, pages 71768–71791. Curran Associates, Inc.