Model Mixture: Merging Task-Specific Language Models

Stanford CS224N Custom Project

Sherry Xie
Department of Computer Science
Stanford University
ycxie@stanford.edu

Elizabeth Zhu
Department of Computer Science
Stanford University
elizhu@stanford.edu

Abstract

Model merging, a method which combines the weights of multiple finetuned models, has emerged as a new technique to build and update foundation models to harbor multi-task capabilities. However, current research in the field focuses primarily on applying model merging to vision models and encoder-decoder models. Our paper explores and evaluates multiple model merging techniques for decoder-only language models. Specifically, we finetune GPT-2 models on three datasets from the GLUE Benchmark: SST-2, COLA, and MRPC, then merge them using linear interpolation and more advanced merging techniques such as TIES-Merging. We find that model merging generally works well for decoder-only language models, but merging performance is affected by dataset characteristics. Specifically, we finetune on specific model layers, finding that finetuning the last layer results in less merging conflict. Finally, we study the effects of scaling up language models on model merging, discovering that larger models improve merging performance.

[Mentor: Yuhui Zhang]

1 Introduction

The standard approach to training a Large Language Model (LLM) to achieve multi-task performance involves curating a vast dataset and requires intensive computational resources to train. As a result, *model merging* (Li et al., 2023) has emerged as a technique to combine the abilities of multiple models at a low cost, enabling those without the necessary compute resources to harness the capabilities of multi-task models while enabling LLMs to be conveniently updated with new task capabilities.

Model merging refers to the process of merging different models by finding optimal ways of combining their parameters. In our paper, we select the decoder-only DistilGPT-2 model as our baseline pretrained model. After finetuning this baseline model on three GLUE benchmark tasks: SST2 (sentiment classification), COLA (linguistic acceptability), and MRPC (paraphase detection), we then conducted model merging on each pairing of these models. We first linearly interpolate the weights of each pair of finetuned models (Choshen et al., 2022; Ilharco et al., 2022). In doing so, we observe that selecting the optimal α during linear interpolation allows the merged model to achieve a merging accuracy of up to 97% when merging models finetuned on similarly structured tasks (SST2 & COLA). In contrast, we observed much higher merging conflicts on pairings with the MRPC dataset, a dataset structured differently than SST2 and COLA as it compares two sentences rather than evaluating one.

We then implement TIES-Merging, an existing and recently proposed merging technique that claims to address parameter interference (Yadav et al., 2023). In a similar fashion, we use DistilGPT-2 as our base pretrained model and merge pairs of models finetuned on different sub-tasks in the GLUE benchmark. However, we discover that TIES-Merging does not perform as well as linear interpolation, perhaps due to the fact that TIES-Merging is specific to encoder-decoder models in addition to our base pretrained model having far fewer parameters than the model TIES-Merging evaluated on.

In addition, we explore whether different pre-merging conditions of the pretrained and finetuned models have an effect on merging performance. We consider two conditions. First, we selectively choose specific *layers* in the pretrained DistilGPT-2 model to finetune, evaluating each pair of finetuned models on the GLUE benchmark after only finetuning their first, middle, or last layer and freezing all other layers. Second, we evaluate whether the size of the pretrained model influences merging performance by conducting linear interpolation experiments with larger models (GPT-2 small, GPT-2 medium) as our base pretrained model. We discover that finetuning on later layers leads to less merging conflict than finetuning on earlier layers and significantly improves the *maximum merging accuracy* for tasks which are structurally different in contrast to when all model layers were finetuned. Finally, we find that models tend to merge better if they are both finetuned on a larger pretrained model, with GPT-2 Medium outperforming DistilGPT-2 in average merging accuracy.

2 Related Work

Model merging is a new and emerging field in NLP. When the field first arrived, there was abundant research on simple averaging techniques to merge model weights (Choshen et al., 2022). Since then, there has been more research on developing more advanced and accurate merging techniques, specifically the TIES-Merging and Fisher Merging techniques (Ilharco et al., 2023; Matena and Raffel, 2022; Yadav et al., 2023). However, most research in model merging has focused on merging vision or vision-language models (Ilharco et al., 2022) and exploring only encoder-decoder models.

We believe that our project makes two major contributions to this existing body of work:

1. We explore the effect of different model-merging techniques on decoder-only models (GPT-2), which are the most state-of-the-art and prevalent language models as of now.
2. We tackle unexplored research questions on how pre-merging conditions such as finetuning specific model layers and scaling the pretrained model influence merging performance. This is an important contribution as current papers tend to focus solely on the merging technique, yet determining the optimal conditions can exert an equal influence on merging performance.

3 Approach

Given our baseline model (DistilGPT-2), we first finetuned the pretrained model separately on three downstream tasks (SST-2, COLA, and MRPC) from the GLUE dataset. We used instruction finetuning (Chung et al., 2022) for our finetuning process, namely, we finetuned language models on tasks phrased as instructions, which enables them to respond better to instructions. We use instruction finetuning because it preserves the model structure and parameter size during model merging. In the instruction finetuning process, we used the following three prompts for each downstream task:

1. SST-2 Prompt: *"{s} This does suggest that this is"* ("good"/"bad")
2. COLA Prompt: *"This sentence {s} is linguistically"* ("acceptable"/"nonacceptable")
3. MRPC Prompt: *"The semantic meanings of '{s1}' and '{s2}' are"* ("same"/"different")

Although we believe the above prompts could be refined (i.e. to be more specific / to flow more naturally), we observed a minimal difference in performance when optimizing the prompt. However, one possible future direction is to explicitly include the options available to the model for the next word, preventing it from generating synonyms that are correct semantically but lower the accuracy.

We choose the above prompt for SST-2 because in evaluation, it achieved a slightly higher accuracy (around 1%) than the more intuitive prompt: *"The sentiment of this sentence {s} is ("positive"/"negative")"*. Furthermore, we chose the above prompt for MRPC because the two words "same" and "different" are directly antonyms but in order to place them in the same sentence, we chose to sacrifice the grammatical correctness of their usage. However, we ensured that all our finetuning accuracies were on par with research benchmarks before proceeding to the merging stage.

The first model merging approach we used is linear-interpolation (Choshen et al., 2022). We define our linear-interpolation process as the following: Suppose we have two sets of parameter weights P and Q from two downstream natural language tasks. We then calculate the final set of parameters R

through the following equation:

$$R_i = \alpha P_i + (1 - \alpha) Q_i \tag{1}$$

For each pairing of the three tasks above, we plugged in 11 α values ranging from 0.0 to 1.0 with an interval of 0.1. The above finetuning and linear-interpolation process is summarized in Figure 1.

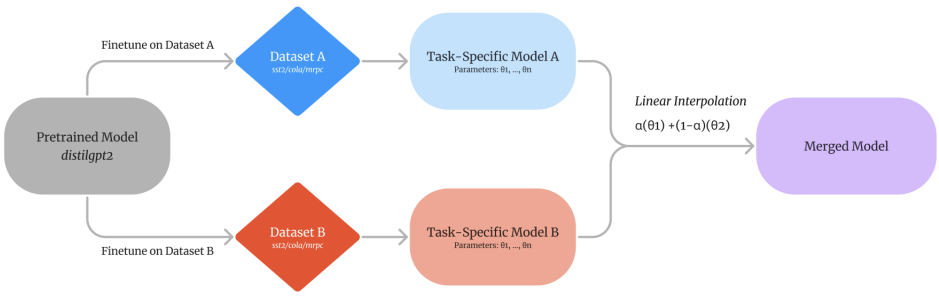

Figure 1: *Merging Finetuned Models with Linear Interpolation*
Given a pretrained DistilGPT-2 model, we finetune it on two different datasets (A and B) respectively and save their parameters. We then use *Equation 1* to linearly interpolate the parameters to produce merged parameters and load them into the pretrained model to obtain the final merged model.

Treating the linear interpolation results as a benchmark, we proceed to implement the TIES-Merging approach (Yadav et al., 2023) for model-merging. This technique uses task vectors (Ilharco et al., 2023), which are created by subtracting the weights of pretrained model from the weights of the task-specific model, after being finetuned on a task. TIES-Merging seeks to address interference in model-merging, namely the idea that parameter signs and magnitudes across models may interfere with each other to the extent of harming model performance. TIES-Merging adopts the following three-step pipeline on task vectors before merging them back with the original pretrained parameters:

1. **Trim**: For each task t, trim a certain percentage of parameters that are deemed to be redundant and set them to 0.

2. **Elect Sign**: Then, create an elected sign vector to resolve disagreements in sign of each vector by only keeping one direction for each parameter.

3. **Disjoint Merge**: Afterwards, compute a *disjoint mean* by only keeping parameter values from models whose signs are aligned with the elected sign selected in the previous step and calculate their mean.

Based on the idea that the performance of these models depends on a small fraction ($k\%$) of parameters, we experimented with different k values ranging from 0.0 to 1.0 with an interval of 0.1 for each pairing of the three downstream tasks mentioned above.

Then, we proceeded to adjust pre-merging conditions. Prior to merging, we first picked three specific layers to finetune (the first layer, the middle layer, and the last layer) on each of the three datasets above, while leaving the other layers untouched. We chose these specific layers with the understanding that earlier model layers capture more general information about the task while later layers capture more task-specific information.

Afterwards, we explored the effect of using a larger pretrained model (GPT-2 Small, GPT-2 Medium) on merging performance, hoping to analyze whether scaling laws would still hold in model merging.

These experiments aim at understanding how setting certain optimal pre-merging conditions can influence model merging results. After the finetuning process, we again used linear interpolation to merge different combinations of finetuned, task-specific models and evaluate their performance.

4 Experiments and Results

4.1 Data

All of our data comes from the GLUE dataset. Specifically, we chose three tasks from this dataset: SST-2, COLA, and MRPC. We chose these three datasets considering our compute capacity, the prevalence of the GLUE benchmark, and the widespread use of these tasks.

With these three datasets, we evaluated our finetuned models on the following three tasks:

1. Classifying the sentiment of a sentence. (SST-2)
2. Identifying linguistic acceptability of sentences. (COLA)
3. Categorizing whether two sentences are paraphrases. (MRPC)

4.2 Evaluation Metrics

We evaluate the performance of our models by testing how it performs on the *validation* dataset specific to each task. Consistent with the instruction finetuning method we used in training, we provided our models with the prompts we trained the models on, while omitting the last word for the model to predict. The evaluation accuracy results we get from running these tests represent the percentage of words correctly predicted by the trained model, out of the validation set.

We define the *merging accuracy* to be the relative performance of our merged model on the two downstream tasks we evaluate it on (see below for a mathematical formalization). A high merging accuracy indicates that our model is capable of performing well on both finetuned tasks, meaning the model can be said to have true multi-task abilities. We calculate the merging accuracy in the following way. Suppose the baseline accuracy from the finetuned models for the two tasks are respectively x and y, and the new accuracy after merging are x' and y', the merging accuracy δ will be:

$$\delta = \frac{1}{2}(\frac{x'}{x} + \frac{y'}{y}) \tag{2}$$

Moreover, for each pair of finetuned models, we perform merging using a range of alpha values from 0.0 to 1.0 with an interval of 0.1. For each merge with a unique alpha value, we evaluate the accuracy of the merged model on both tasks individually, and calculate the merging accuracy as defined above. For each merged model emerging from a pair of task-specific models, we specifically consider two important metrics: a) the *maximum* merging accuracy across all the alpha values and b) the *average* merging accuracy across all the alpha values. For our purposes, the maximum merging accuracy is a more important evaluation metric to optimize for as, in the real world, we are primarily interested in the best merging results and notably, we are able to select the optimal alpha to produce those results.

4.3 Linear Interpolation

In these sets of experiments, we linearly interpolated all pairs of downstream tasks we selected (SST2 + COLA, SST2 + MRPC, COLA + MRPC) using 11 different α values as mentioned above.

We observed that both maximum and average merging accuracies tend to be significantly higher when merging models finetuned on SST2 + COLA in comparison to SST2 + MRPC and COLA + MRPC. This shows that linear interpolation generally functions well as a model merging technique, but task/dataset characteristics determine how high the merging accuracy between a given pair is.

4.4 TIES Merging

After the linear interpolation stage, we also did pairwise parameter merging on the same tasks (SST2 + COLA, SST2 + MRPC, COLA + MRPC) using the TIES-Merging approach. We tuned the hyperparameter named k in the paper, which determines the percentage of high-magnitude task vectors we want to keep in the merging process. Because we neither wanted to discard all parameters nor keep all parameters, we attempted k values between 0.1 and 0.9 with 0.1 as the increment.

We found that TIES-Merging had far worse merging accuracy in comparison to linear interpolation. This could be because we used a decoder-only, 80 million parameter pretrained model, which greatly differs from the encoder-decoder, 220 million parameter model used in the TIES-Merging paper.

4.5 Finetuning Specific Layers

Contrary to our expectations, for finetuned models which failed to merge well through linear interpolation (specifically, the pairs of SST2 + MRPC and COLA + MRPC), we achieved better results when we merged models finetuned only on the last layer of the pretrained models rather than finetuned on all layers. When we originally merged models that were finetuned on COLA and MRPC respectively on all layers, the maximum merging accuracy we achieved was 52.93%, yet when finetuning these models only on the last layer, the maximum merging accuracy was 66.51%. In addition, the average merging accuracy was roughly 10% better when finetuning only the last layer. Merging models finetuned solely on their first layer and their middle layer resulted in bad overall merging performance.

4.6 Scaling the Size of the Pretrained Model

When we scaled the size of the pretrained model we used to finetune on different tasks, namely from DistilGPT-2 to larger models such as GPT-2 small or GPT-2 medium, we observed overall better merging performance when using the larger models. Specifically, when substituting DistilGPT-2 with GPT-2 medium for our pretrained model, we not only get higher individual accuracies on both datasets prior to merging but we also see a significant increase in both maximum and average merging accuracy when merging tasks that are not structurally similar. Notably, models finetuned on COLA and MRPC achieve a 20% increase in average model accuracy compared to using DistilGPT-2. It is worth noting that GPT-2 small achieves comparable, albeit slightly worse, maximum merging accuracies to DistilGPT-2 on SST2 + COLA and SST2 + MRPC, while significantly improving the merging accuracy for COLA + MRPC. We hypothesize that since DistilGPT-2 is a comparable lightweight version of GPT-2 small, the effects from scaling up our pretrained model are not as noticeable as when we scaled up to GPT-2 medium, which has significantly more parameters.

4.7 Experiment Results Summary

We present the following two tables (Tables 1 & 2) to summarize our maximum merging accuracies and average merging accuracies for all four sets of experiments we conducted.

We selected the maximum merging accuracy as a primary evaluation metric because we generally care about choosing a single best merging α when we conduct model merging, to optimize for the best result. However, we additionally evaluated on the average merging accuracy because the average values reflect in general how effective a certain merging technique or pre-merging condition is.

Experiment \ Merged Tasks	SST2 + COLA	SST2 + MRPC	COLA + MRPC
Linear Merging	97.22%	69.20%	52.93%
TIES Merging	96.78%	49.05%	49.21%
Finetuning First Layer	49.15%	50.88%	50%
Finetuning Middle Layer	50.42%	58.86%	50.90%
Finetuning Last Layer	95.64%	63.93%	66.51%
Pretrain GPT-2 Small	95.26%	65.07%	79.59%
Pretrain GPT-2 Medium	93.53%	80.33%	92.38%

Table 1: Max Merging Accuracy for each experiment

5 Analysis

Our main finding is that weight interpolation is an effective technique to merge decoder-only models. A consistent throughline of our findings is that models finetuned on COLA and SST-2 respectively

Experiment \ Merged Tasks	SST2 + COLA	SST2 + MRPC	COLA + MRPC
Linear Merging	63.38%	54.01%	41.89%
TIES Merging	52.00%	37.66%	37.21%
Finetuning First Layer	33.35%	33.04%	34.56%
Finetuning Middle Layer	47.95%	51.15%	48.18%
Finetuning Last Layer	61.68%	53.00%	51.19%
Pretrain GPT-2 Small	56.72%	53.00%	54.63%
Pretrain GPT-2 Medium	61.21%	58.39%	61.12%

Table 2: Average Merging Accuracy for each experiment

merge much better than the other pairs (namely, COLA and MRPC and SST-2 and MRPC). We hypothesize this is because COLA and SST-2 are tasks that have similarly structured datasets, as both evaluate properties of a single sentence, meaning the models likely share more compatible parameters. From this, we conclude that although model merging generally works well for combining task-specific models, its performance depends heavily on dataset characteristics. As a work around, we found that for tasks with different structural characteristics, finetuning the pretrained model on specific layers and scaling up the pretrained model can improve their merging performance.

5.1 Comparing Merging Methods: Linear Interpolation vs. TIES-Merging

We expected the distribution of merging accuracy results among the range of alpha values we used when conducting a linear interpolation of the model weights. Specifically, it makes sense that the highest merging accuracy occurred with an alpha value of around 0.5 (falling within a range of alpha values between 0.4 to 0.6) as that equally balances the capabilities of both finetuned models, not overly sacrificing performance on one for another.

Moreover, the lower average merging accuracies when merging a model either finetuned on SST2 or COLA with one finetuned on MRPC can likely be attributed to structural differences in the task described by the MRPC dataset as it involves comparing the semantic meaning of *two* sentences as opposed to analyzing simply one sentence, as SST2 and COLA do. As a result, merging models finetuned on SST2 and COLA results in better merging performance as the parameters of both finetuned models are more likely to capture compatible features related to the meaning of a single sentence, whereas merging either of these models with a model finetuned on the MRPC task is more prone to parameter interference, preventing the merged model from performing well on both tasks.

Analyzing the two merging techniques we implemented, we did not expect linear interpolation to perform much better than TIES-Merging since we assumed that the more advanced approach would perform better. However, we offer two possible explanations for this, as follows.

One potential explanation for this is that TIES-Merging tailored their merging technique to T-5 base and T-5 large, which are both encoder-decoder transformer models (Vaswani et al., 2023), whereas distilgpt2 belongs to the family of decoder-only models (Radford and Narasimhan, 2018).

A second potential explanation for this may be that TIES-Merging is meant to address parameter interference in merging larger models. T-5 base has 220 million parameters and T-5 large has 770 million parameters, whereas DistilGPT-2, the pretrained model we used, has a mere 82 million parameters. Since we used a pretrained model with much fewer parameters, we hypothesize that our merging process implicitly results in fewer instances of parameter interference, making TIES-Merging not a suitable approach for our use case.

5.2 Effect of Finetuning Specific Layers on Merging Performance

We observed that finetuning only the last layer of the pretrained model on a specific task resulted in improved merging performance in comparison to finetuning only the first or middle layer. In addition, for the pairs of finetuned models which failed to merge well during linear interpolation, finetuning them on solely the last layer resulted in better merging performance than when we finetuned them on

all layers. Moreover, in addition to achieving comparable or better merging performance depending on the dataset, finetuning on only the last layer is evidently more computationally efficient and requires interpolating fewer weights, making it an appealing condition for merging dissimilar tasks.

We hypothesize that this is because in finetuning only the last layer of the pretrained model and freezing the other layers, we maintain parameters in the earlier layers which encode the more general capabilities of the model. In doing so, we minimize the risk of the model overfitting to any particular finetuned task, thereby facilitating merging of its weights with another task-specific model. Preserving general abilities is particularly important for merging models finetuned on MRPC with both SST2 and COLA since MRPC is a structurally different task from SST2 and COLA as we explained earlier, meaning it benefits more from preserving parameters that capture general capabilities.

5.3 Effect of Scaling the Pretrained Model on Merging Performance

We observed that larger models such as GPT2-Medium result in better merging accuracies than smaller models. We hypothesize this is because a model with a greater number of parameters better generalizes to a variety of tasks, improving the ability for two task-specific models to merge well.

6 Conclusion

In this paper, we contribute to the emerging field of model merging by investigating various merging techniques and specific pre-merging conditions that can improve merging performance. Our paper offers an important contribution to the field of NLP as our paper offers guidance on a potentially optimal set of merging techniques and model traits that can enable models to effectively acquire new skills with no additional training, and achieve decent performance in two entirely different tasks.

Specifically, our paper demonstrates three primary findings: that linear interpolation is effective at preserving the merged model's performance on both datasets, finetuning on only the last layer of the pretrained model results in better merging accuracy on pairs of models finetuned on structurally dissimilar tasks, and using a larger pretrained model results in better merging performance.

A limitation of our work is that we were constrained to experimenting with relatively small pretrained models due to computational constraints, meaning that we are unable to verify if these results are consistent with larger models. In the future, we would like to experiment with merging three or more finetuned models to achieve enhanced multi-task capabilities and further scaling the models at hand.

References

Leshem Choshen, Elad Venezian, Noam Slonim, and Yoav Katz. 2022. Fusing finetuned models for better pretraining.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic.

Gabriel Ilharco, Mitchell Wortsman, Samir Yitzhak Gadre, Shuran Song, Hannaneh Hajishirzi, Simon Kornblith, Ali Farhadi, and Ludwig Schmidt. 2022. Patching open-vocabulary models by interpolating weights.

Weishi Li, Yong Peng, Miao Zhang, Liang Ding, Han Hu, and Li Shen. 2023. Deep model fusion: A survey.

Michael Matena and Colin Raffel. 2022. Merging models with fisher-weighted averaging.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. 2023. Ties-merging: Resolving interference when merging models.

A Appendix

Here, we share all of our merging results for every single experiment we ran.

Model	Alpha	SST2	COLA	Merging Performance
0.0 Fine-tuned model on SST2 + 1.0 Fine-tuned model on COLA	0	0.00%	69.03%	50%
0.1 Fine-tuned model on SST2 + 0.9 Fine-tuned model on COLA	0.1	0.00%	69.03%	50%
0.2 Fine-tuned model on SST2 + 0.8 Fine-tuned model on COLA	0.2	0.57%	68.94%	50%
0.3 Fine-tuned model on SST2 + 0.7 Fine-tuned model on COLA	0.3	2.18%	68.74%	51%
0.4 Fine-tuned model on SST2 + 0.6 Fine-tuned model on COLA	0.4	26.61%	69.03%	65%
0.5 Fine-tuned model on SST2 + 0.5 Fine-tuned model on COLA	0.5	71.10%	69.03%	90%
0.6 Fine-tuned model on SST2 + 0.4 Fine-tuned model on COLA	0.6	83.03%	68.84%	97.22%
0.7 Fine-tuned model on SST2 + 0.3 Fine-tuned model on COLA	0.7	85.09%	60.69%	92%
0.8 Fine-tuned model on SST2 + 0.2 Fine-tuned model on COLA	0.8	86.35%	3.45%	52%
0.9 Fine-tuned model on SST2 + 0.1 Fine-tuned model on COLA	0.9	87.39%	0.00%	50%
1.0 Fine-tuned model on SST2 + 0.0 Fine-tuned model on COLA	1.0	87.96%	0.00%	50%

Figure 2: Linear Interpolation: SST2 + COLA

Model	Alpha	SST2	MRPC	Merging Performance
0.0 Fine-tuned model on SST2 + 1.0 Fine-tuned model on MRPC	0	0.00%	77.70%	50%
0.1 Fine-tuned model on SST2 + 0.9 Fine-tuned model on MRPC	0.1	0.00%	78.19%	50.30%
0.2 Fine-tuned model on SST2 + 0.8 Fine-tuned model on MRPC	0.2	0.11%	75.98%	48.96%
0.3 Fine-tuned model on SST2 + 0.7 Fine-tuned model on MRPC	0.3	0.69%	76.23%	49.44%
0.4 Fine-tuned model on SST2 + 0.6 Fine-tuned model on MRPC	0.4	18.92%	68.14%	54.60%
0.5 Fine-tuned model on SST2 + 0.5 Fine-tuned model on MRPC	0.5	63.19%	51.72%	69.20%
0.6 Fine-tuned model on SST2 + 0.4 Fine-tuned model on MRPC	0.6	80.05%	31.62%	65.85%
0.7 Fine-tuned model on SST2 + 0.3 Fine-tuned model on MRPC	0.7	84.17%	1.96%	49.10%
0.8 Fine-tuned model on SST2 + 0.2 Fine-tuned model on MRPC	0.8	86.01%	0.00%	48.90%
0.9 Fine-tuned model on SST2 + 0.1 Fine-tuned model on MRPC	0.9	87.50%	0.00%	49.74%
1.0 Fine-tuned model on SST2 + 0.0 Fine-tuned model on MRPC	1.0	93.69%	0.00%	50%

Figure 3: Linear Interpolation: SST2 + MRPC

Model	Alpha	COLA	MRPC	Merging Performance
0.0 Fine-tuned model on COLA + 1.0 Fine-tuned model on MRPC	0	0.00%	77.70%	50%
0.1 Fine-tuned model on COLA + 0.9 Fine-tuned model on MRPC	0.1	0.00%	78.19%	50.32%
0.2 Fine-tuned model on COLA + 0.8 Fine-tuned model on MRPC	0.2	0.00%	76.47%	49.21%
0.3 Fine-tuned model on COLA + 0.7 Fine-tuned model on MRPC	0.3	0.00%	76.72%	49.37%
0.4 Fine-tuned model on COLA + 0.6 Fine-tuned model on MRPC	0.4	0.86%	74.51%	48.57%
0.5 Fine-tuned model on COLA + 0.5 Fine-tuned model on MRPC	0.5	16.59%	62.75%	52.40%
0.6 Fine-tuned model on COLA + 0.4 Fine-tuned model on MRPC	0.6	30.39%	48.04%	52.93%
0.7 Fine-tuned model on COLA + 0.3 Fine-tuned model on MRPC	0.7	31.83%	11.27%	30.31%
0.8 Fine-tuned model on COLA + 0.2 Fine-tuned model on MRPC	0.8	40.65%	0.00%	29.44%
0.9 Fine-tuned model on COLA + 0.1 Fine-tuned model on MRPC	0.9	66.63%	0.00%	48.26%
1.0 Fine-tuned model on COLA + 0.0 Fine-tuned model on MRPC	1.0	69.03%	0.00%	50%

Figure 4: Linear Interpolation: COLA + MRPC

K	SST2	COLA	Merging Performance
0.1	0.00%	0.00%	0.00%
0.2	0.23%	8.15%	6.03%
0.3	38.53%	68.74%	71.86%
0.4	81.88%	68.94%	96.78%
0.5	84.75%	60.02%	91.96%
0.6	86.35%	4.99%	53.02%
0.7	86.12%	0.00%	49.27%
0.8	85.89%	0.00%	49.14%
0.9	87.39%	0.00%	50%

Figure 5: TIES Merging: SST2 + COLA

K	SST2	MRPC	Merging Performance
0.1	0.00%	0.00%	0.00%
0.2	0.11%	12.75%	8.27%
0.3	5.50%	57.60%	40.19%
0.4	1.95%	71.32%	47.00%
0.5	0.23%	74.75%	48.23%
0.6	0.00%	75.49%	48.58%
0.7	0.00%	76.23%	49.05%
0.8	0.00%	76.47%	49.21%
0.9	0.00%	75.25%	48.42%

Figure 6: TIES Merging: SST2 + MRPC

K	COLA	MRPC	Merging Performance
0.1	0.00%	0.00%	0.00%
0.2	0.00%	11.03%	7.10%
0.3	0.00%	62.75%	39.93%
0.4	0.00%	72.79%	46.84%
0.5	0.00%	73.28%	47.22%
0.6	0.00%	74.51%	47.94%
0.7	0.00%	76.47%	49.21%
0.8	0.00%	76.23%	49.05%
0.9	0.00%	74.02%	47.63%

Figure 7: TIES Merging: COLA + MRPC

Model	Alpha	SST2	COLA	Merging Performance
0.0 Fine-tuned model on SST2 + 1.0 Fine-tuned model on COLA	0	0.00%	69.13%	50%
0.1 Fine-tuned model on SST2 + 0.9 Fine-tuned model on COLA	0.1	0.00%	68.94%	49.86%
0.2 Fine-tuned model on SST2 + 0.8 Fine-tuned model on COLA	0.2	0.00%	68.07%	49.23%
0.3 Fine-tuned model on SST2 + 0.7 Fine-tuned model on COLA	0.3	0.23%	67.50%	48.95%
0.4 Fine-tuned model on SST2 + 0.6 Fine-tuned model on COLA	0.4	2.52%	67.50%	50.21%
0.5 Fine-tuned model on SST2 + 0.5 Fine-tuned model on COLA	0.5	54.01%	68.26%	79.22%
0.6 Fine-tuned model on SST2 + 0.4 Fine-tuned model on COLA	0.6	87.04%	65.20%	95.26%
0.7 Fine-tuned model on SST2 + 0.3 Fine-tuned model on COLA	0.7	88.99%	3.07%	51.40%
0.8 Fine-tuned model on SST2 + 0.2 Fine-tuned model on COLA	0.8	90.02%	0.00%	49.75%
0.9 Fine-tuned model on SST2 + 0.1 Fine-tuned model on COLA	0.9	90.60%	0.00%	50.07%
1.0 Fine-tuned model on SST2 + 0.0 Fine-tuned model on COLA	1	90.48%	0.00%	50%

Figure 8: GPT-2 Small: SST2 + COLA

Model	Alpha	SST2	MRPC	Merging Performance
0.0 Fine-tuned model on SST2 + 1.0 Fine-tuned model on MRPC	0	0.00%	74.26%	50%
0.1 Fine-tuned model on SST2 + 0.9 Fine-tuned model on MRPC	0.1	0.00%	75.00%	50.50%
0.2 Fine-tuned model on SST2 + 0.8 Fine-tuned model on MRPC	0.2	0.11%	76.23%	51.39%
0.3 Fine-tuned model on SST2 + 0.7 Fine-tuned model on MRPC	0.3	0.57%	74.02%	50.15%
0.4 Fine-tuned model on SST2 + 0.6 Fine-tuned model on MRPC	0.4	8.83%	65.20%	48.78%
0.5 Fine-tuned model on SST2 + 0.5 Fine-tuned model on MRPC	0.5	49.31%	41.91%	55.47%
0.6 Fine-tuned model on SST2 + 0.4 Fine-tuned model on MRPC	0.6	80.73%	30.39%	65.07%
0.7 Fine-tuned model on SST2 + 0.3 Fine-tuned model on MRPC	0.7	88.30%	19.85%	62/16%
0.8 Fine-tuned model on SST2 + 0.2 Fine-tuned model on MRPC	0.8	89.79%	0.00%	49.62%
0.9 Fine-tuned model on SST2 + 0.1 Fine-tuned model on MRPC	0.9	90.37%	0.00%	49.94%
1.0 Fine-tuned model on SST2 + 0.0 Fine-tuned model on MRPC	1	90.48%	0.00%	50%

Figure 9: GPT-2 Small: SST2 + MRPC

Model	Alpha	COLA	MRPC	Merging Performance
0.0 Fine-tuned model on COLA + 1.0 Fine-tuned model on MRPC	0	0.00%	74.26%	50%
0.1 Fine-tuned model on COLA + 0.9 Fine-tuned model on MRPC	0.1	0.00%	74.51%	50.17%
0.2 Fine-tuned model on COLA + 0.8 Fine-tuned model on MRPC	0.2	0.00%	75.74%	51.00%
0.3 Fine-tuned model on COLA + 0.7 Fine-tuned model on MRPC	0.3	0.00%	79.17%	53.31%
0.4 Fine-tuned model on COLA + 0.6 Fine-tuned model on MRPC	0.4	0.00%	73.28%	49.34%
0.5 Fine-tuned model on COLA + 0.5 Fine-tuned model on MRPC	0.5	2.68%	65.20%	45.84%
0.6 Fine-tuned model on COLA + 0.4 Fine-tuned model on MRPC	0.6	63.95%	49.51%	79.59%
0.7 Fine-tuned model on COLA + 0.3 Fine-tuned model on MRPC	0.7	66.35%	36.03%	72.25%
0.8 Fine-tuned model on COLA + 0.2 Fine-tuned model on MRPC	0.8	68.74%	0.00%	49.72%
0.9 Fine-tuned model on COLA + 0.1 Fine-tuned model on MRPC	0.9	68.74%	0.00%	49.72%
1.0 Fine-tuned model on COLA + 0.0 Fine-tuned model on MRPC	1	69.13%	0.00%	50%

Figure 10: GPT-2 Small: COLA + MRPC

Model	Alpha	SST2	COLA	Merging Performance
0.0 Fine-tuned model on SST2 + 1.0 Fine-tuned model on COLA	0	0.00%	69.13%	50%
0.1 Fine-tuned model on SST2 + 0.9 Fine-tuned model on COLA	0.1	0.11%	69.13%	50.05%
0.2 Fine-tuned model on SST2 + 0.8 Fine-tuned model on COLA	0.2	0.34%	69.13%	50.18%
0.3 Fine-tuned model on SST2 + 0.7 Fine-tuned model on COLA	0.3	7.34%	69.13%	53.91%
0.4 Fine-tuned model on SST2 + 0.6 Fine-tuned model on COLA	0.4	55.62%	68.74%	79.40%
0.5 Fine-tuned model on SST2 + 0.5 Fine-tuned model on COLA	0.5	84.17%	67.21%	93.53%
0.6 Fine-tuned model on SST2 + 0.4 Fine-tuned model on COLA	0.6	89.11%	60.02%	90.67%
0.7 Fine-tuned model on SST2 + 0.3 Fine-tuned model on COLA	0.7	91.51%	10.26%	56.26%
0.8 Fine-tuned model on SST2 + 0.2 Fine-tuned model on COLA	0.8	92.66%	0.00%	49.45%
0.9 Fine-tuned model on SST2 + 0.1 Fine-tuned model on COLA	0.9	93.46%	0.00%	49.88%
1.0 Fine-tuned model on SST2 + 0.0 Fine-tuned model on COLA	1	93.69%	0.00%	50%

Figure 11: GPT-2 Medium: SST2 + COLA

Model	Alpha	SST2	MRPC	Merging Performance
0.0 Fine-tuned model on SST2 + 1.0 Fine-tuned model on MRPC	0	0.00%	81.13%	50%
0.1 Fine-tuned model on SST2 + 0.9 Fine-tuned model on MRPC	0.1	0.00%	80.64%	49.70%
0.2 Fine-tuned model on SST2 + 0.8 Fine-tuned model on MRPC	0.2	0.00%	80.15%	49.40%
0.3 Fine-tuned model on SST2 + 0.7 Fine-tuned model on MRPC	0.3	0.23%	79.17%	48.91%
0.4 Fine-tuned model on SST2 + 0.6 Fine-tuned model on MRPC	0.4	14.56%	78.92%	56.41%
0.5 Fine-tuned model on SST2 + 0.5 Fine-tuned model on MRPC	0.5	66.17%	73.04%	80.33%
0.6 Fine-tuned model on SST2 + 0.4 Fine-tuned model on MRPC	0.6	89.11%	43.63%	74.44%
0.7 Fine-tuned model on SST2 + 0.3 Fine-tuned model on MRPC	0.7	91.86%	31.37%	68.36%
0.8 Fine-tuned model on SST2 + 0.2 Fine-tuned model on MRPC	0.8	92.43%	25.25%	64.89%
0.9 Fine-tuned model on SST2 + 0.1 Fine-tuned model on MRPC	0.9	93.35%	0.00%	49.82%
1.0 Fine-tuned model on SST2 + 0.0 Fine-tuned model on MRPC	1	93.69%	0.00%	50%

Figure 12: GPT-2 Medium: SST2 + MRPC

Model	Alpha	COLA	MRPC	Merging Performance
0.0 Fine-tuned model on COLA + 1.0 Fine-tuned model on MRPC	0	0.00%	81.13%	50%
0.1 Fine-tuned model on COLA + 0.9 Fine-tuned model on MRPC	0.1	0.00%	80.15%	49.40%
0.2 Fine-tuned model on COLA + 0.8 Fine-tuned model on MRPC	0.2	0.10%	80.15%	49.47%
0.3 Fine-tuned model on COLA + 0.7 Fine-tuned model on MRPC	0.3	0.38%	79.17%	49.07%
0.4 Fine-tuned model on COLA + 0.6 Fine-tuned model on MRPC	0.4	9.78%	78.68%	55.56%
0.5 Fine-tuned model on COLA + 0.5 Fine-tuned model on MRPC	0.5	51.97%	77.70%	85.47%
0.6 Fine-tuned model on COLA + 0.4 Fine-tuned model on MRPC	0.6	64.24%	74.51%	92.38%
0.7 Fine-tuned model on COLA + 0.3 Fine-tuned model on MRPC	0.7	67.69%	38.24%	72.52%
0.8 Fine-tuned model on COLA + 0.2 Fine-tuned model on MRPC	0.8	68.94%	29.90%	68.29%
0.9 Fine-tuned model on COLA + 0.1 Fine-tuned model on MRPC	0.9	69.13%	0.25%	50.15%
1.0 Fine-tuned model on COLA + 0.0 Fine-tuned model on MRPC	1	69.13%	0.00%	50%

Figure 13: GPT-2 Medium: COLA + MRPC

Model	Alpha	SST2	COLA	Merging Performance
0.0 Fine-tuned model on SST2 + 1.0 Fine-tuned model on COLA	0	0.00%	67.59%	50%
0.1 Fine-tuned model on SST2 + 0.9 Fine-tuned model on COLA	0.1	0.00%	66.44%	49.15%
0.2 Fine-tuned model on SST2 + 0.8 Fine-tuned model on COLA	0.2	0.00%	65.10%	45.16%
0.3 Fine-tuned model on SST2 + 0.7 Fine-tuned model on COLA	0.3	0.00%	42.57%	31.49%
0.4 Fine-tuned model on SST2 + 0.6 Fine-tuned model on COLA	0.4	0.11%	1.25%	1.01%
0.5 Fine-tuned model on SST2 + 0.5 Fine-tuned model on COLA	0.5	0.57%	0.00%	0.45%
0.6 Fine-tuned model on SST2 + 0.4 Fine-tuned model on COLA	0.6	10.32%	0.00%	8.12%
0.7 Fine-tuned model on SST2 + 0.3 Fine-tuned model on COLA	0.7	51.72%	0.00%	40.71%
0.8 Fine-tuned model on SST2 + 0.2 Fine-tuned model on COLA	0.8	56.88%	0.00%	44.77%
0.9 Fine-tuned model on SST2 + 0.1 Fine-tuned model on COLA	0.9	58.49%	0.00%	46.03%
1.0 Fine-tuned model on SST2 + 0.0 Fine-tuned model on COLA	1.0	63.53%	0.00%	50%

Figure 14: Finetune First Layer: SST2 + COLA

Model	Alpha	SST2	COLA	Merging Performance
0.0 Fine-tuned model on SST2 + 1.0 Fine-tuned model on COLA	0	0.00%	67.88%	50%
0.1 Fine-tuned model on SST2 + 0.9 Fine-tuned model on COLA	0.1	0.00%	67.59%	49.79%
0.2 Fine-tuned model on SST2 + 0.8 Fine-tuned model on COLA	0.2	0.00%	67.21%	49.51%
0.3 Fine-tuned model on SST2 + 0.7 Fine-tuned model on COLA	0.3	0.00%	63.47%	46.75%
0.4 Fine-tuned model on SST2 + 0.6 Fine-tuned model on COLA	0.4	4.59%	48.42%	38.42%
0.5 Fine-tuned model on SST2 + 0.5 Fine-tuned model on COLA	0.5	52.52%	15.63%	43.04%
0.6 Fine-tuned model on SST2 + 0.4 Fine-tuned model on COLA	0.6	80.62%	2.11%	50.10%
0.7 Fine-tuned model on SST2 + 0.3 Fine-tuned model on COLA	0.7	82.22%	0.00%	49.51%
0.8 Fine-tuned model on SST2 + 0.2 Fine-tuned model on COLA	0.8	82.91%	0.00%	49.93%
0.9 Fine-tuned model on SST2 + 0.1 Fine-tuned model on COLA	0.9	83.72%	0.00%	50.42%
1.0 Fine-tuned model on SST2 + 0.0 Fine-tuned model on COLA	1.0	83.03%	0.00%	50%

Figure 15: Finetune Middle Layer: SST2 + COLA

Model	Alpha	SST2	COLA	Merging Performance
0.0 Fine-tuned model on SST2 + 1.0 Fine-tuned model on COLA	0	0.00%	50.34%	50%
0.1 Fine-tuned model on SST2 + 0.9 Fine-tuned model on COLA	0.1	0.00%	55.61%	55.23%
0.2 Fine-tuned model on SST2 + 0.8 Fine-tuned model on COLA	0.2	0.57%	61.74	61.69%
0.3 Fine-tuned model on SST2 + 0.7 Fine-tuned model on COLA	0.3	8.37%	63.57%	68.55%
0.4 Fine-tuned model on SST2 + 0.6 Fine-tuned model on COLA	0.4	35.09%	65.48%	87.70%
0.5 Fine-tuned model on SST2 + 0.5 Fine-tuned model on COLA	0.5	55.62%	60.12%	95.64%
0.6 Fine-tuned model on SST2 + 0.4 Fine-tuned model on COLA	0.6	61.62%	30.39%	69.99%
0.7 Fine-tuned model on SST2 + 0.3 Fine-tuned model on COLA	0.7	65.48%	3.16%	45.43%
0.8 Fine-tuned model on SST2 + 0.2 Fine-tuned model on COLA	0.8	70.76%	0.00%	45.70%
0.9 Fine-tuned model on SST2 + 0.1 Fine-tuned model on COLA	0.9	75.11%	0.00%	48.51%
1.0 Fine-tuned model on SST2 + 0.0 Fine-tuned model on COLA	1.0	77.41%	0.00%	50%

Figure 16: Finetune Last Layer: SST2 + COLA

Model	Alpha	SST2	MRPC	Merging Performance
0.0 Fine-tuned model on SST2 + 1.0 Fine-tuned model on MRPC	0	0.00%	69.85%	50%
0.1 Fine-tuned model on SST2 + 0.9 Fine-tuned model on MRPC	0.1	0.11%	71.08%	50.88%
0.2 Fine-tuned model on SST2 + 0.8 Fine-tuned model on MRPC	0.2	0.11%	38.24%	27.46%
0.3 Fine-tuned model on SST2 + 0.7 Fine-tuned model on MRPC	0.3	0.11%	31.62%	22.72%
0.4 Fine-tuned model on SST2 + 0.6 Fine-tuned model on MRPC	0.4	0.46%	29.66%	21.59%
0.5 Fine-tuned model on SST2 + 0.5 Fine-tuned model on MRPC	0.5	0.92%	0.98%	1.43%
0.6 Fine-tuned model on SST2 + 0.4 Fine-tuned model on MRPC	0.6	8.26%	0.00%	6.50%
0.7 Fine-tuned model on SST2 + 0.3 Fine-tuned model on MRPC	0.7	47.25%	0.00%	37.19%
0.8 Fine-tuned model on SST2 + 0.2 Fine-tuned model on MRPC	0.8	61.23%	0.00%	48.19%
0.9 Fine-tuned model on SST2 + 0.1 Fine-tuned model on MRPC	0.9	60.44%	0.00%	47.47%
1.0 Fine-tuned model on SST2 + 0.0 Fine-tuned model on MRPC	1.0	63.53%	0.00%	50%

Figure 17: Finetune First Layer: SST2 + MRPC

Model	Alpha	SST2	MRPC	Merging Performance
0.0 Fine-tuned model on SST2 + 1.0 Fine-tuned model on MRPC	0	0.00%	73.53%	50%
0.1 Fine-tuned model on SST2 + 0.9 Fine-tuned model on MRPC	0.1	0.11%	74.26%	50.56%
0.2 Fine-tuned model on SST2 + 0.8 Fine-tuned model on MRPC	0.2	0.46%	74.51%	50.94%
0.3 Fine-tuned model on SST2 + 0.7 Fine-tuned model on MRPC	0.3	0.80%	74.51%	51.15%
0.4 Fine-tuned model on SST2 + 0.6 Fine-tuned model on MRPC	0.4	6.08%	69.85%	51.16%
0.5 Fine-tuned model on SST2 + 0.5 Fine-tuned model on MRPC	0.5	58.37%	34.80%	58.86%
0.6 Fine-tuned model on SST2 + 0.4 Fine-tuned model on MRPC	0.6	81.77%	0.49%	49.65%
0.7 Fine-tuned model on SST2 + 0.3 Fine-tuned model on MRPC	0.7	82.34%	0.00%	49.66%
0.8 Fine-tuned model on SST2 + 0.2 Fine-tuned model on MRPC	0.8	83.60%	0.00%	50.42%
0.9 Fine-tuned model on SST2 + 0.1 Fine-tuned model on MRPC	0.9	83.37%	0.00%	50.28%
1.0 Fine-tuned model on SST2 + 0.0 Fine-tuned model on MRPC	1.0	82.91%	0.00%	50%

Figure 18: Finetune Middle Layer: SST2 + MRPC

Model	Alpha	SST2	MRPC	Merging Performance
0.0 Fine-tuned model on SST2 + 1.0 Fine-tuned model on MRPC	0	0.00%	73.53%	50%
0.1 Fine-tuned model on SST2 + 0.9 Fine-tuned model on MRPC	0.1	0.00%	73.28%	49.83%
0.2 Fine-tuned model on SST2 + 0.8 Fine-tuned model on MRPC	0.2	0.00%	71.08%	48.33%
0.3 Fine-tuned model on SST2 + 0.7 Fine-tuned model on MRPC	0.3	5.62%	65.20%	47.97%
0.4 Fine-tuned model on SST2 + 0.6 Fine-tuned model on MRPC	0.4	42.20%	53.92%	63.93%
0.5 Fine-tuned model on SST2 + 0.5 Fine-tuned model on MRPC	0.5	71.79%	28.92%	66.04%
0.6 Fine-tuned model on SST2 + 0.4 Fine-tuned model on MRPC	0.6	79.24%	3.92%	53.85%
0.7 Fine-tuned model on SST2 + 0.3 Fine-tuned model on MRPC	0.7	79.47%	0.25%	51.50%
0.8 Fine-tuned model on SST2 + 0.2 Fine-tuned model on MRPC	0.8	79.01%	0.00%	51.03%
0.9 Fine-tuned model on SST2 + 0.1 Fine-tuned model on MRPC	0.9	78.21%	0.00%	50.52%
1.0 Fine-tuned model on SST2 + 0.0 Fine-tuned model on MRPC	1.0	77.41%	0.00%	50%

Figure 19: Finetune Last Layer: SST2 + MRPC

Model	Alpha	COLA	MRPC	Merging Performance
0.0 Fine-tuned model on COLA + 1.0 Fine-tuned model on MRPC	0	0.00%	69.85%	50%
0.1 Fine-tuned model on COLA + 0.9 Fine-tuned model on MRPC	0.1	0.00%	69.12%	50%
0.2 Fine-tuned model on COLA + 0.8 Fine-tuned model on MRPC	0.2	0.10%	44.36%	31.83%
0.3 Fine-tuned model on COLA + 0.7 Fine-tuned model on MRPC	0.3	0.10%	31.62%	22.71%
0.4 Fine-tuned model on COLA + 0.6 Fine-tuned model on MRPC	0.4	0.19%	31.62%	22.77%
0.5 Fine-tuned model on COLA + 0.5 Fine-tuned model on MRPC	0.5	0.77%	10.78%	8.29%
0.6 Fine-tuned model on COLA + 0.4 Fine-tuned model on MRPC	0.6	12.94%	0.00%	9.57%
0.7 Fine-tuned model on COLA + 0.3 Fine-tuned model on MRPC	0.7	55.32%	0.00%	40.92%
0.8 Fine-tuned model on COLA + 0.2 Fine-tuned model on MRPC	0.8	62.32%	0.00%	46.10%
0.9 Fine-tuned model on COLA + 0.1 Fine-tuned model on MRPC	0.9	64.81%	0.00%	47.94%
1.0 Fine-tuned model on COLA + 0.0 Fine-tuned model on MRPC	1.0	67.59%	0.00%	50%

Figure 20: Finetune First Layer: COLA + MRPC

Model	Alpha	COLA	MRPC	Merging Performance
0.0 Fine-tuned model on COLA + 1.0 Fine-tuned model on MRPC	0	0.00%	73.53%	50%
0.1 Fine-tuned model on COLA + 0.9 Fine-tuned model on MRPC	0.1	0.00%	74.26%	50.50%
0.2 Fine-tuned model on COLA + 0.8 Fine-tuned model on MRPC	0.2	0.10%	74.75%	50.90%
0.3 Fine-tuned model on COLA + 0.7 Fine-tuned model on MRPC	0.3	0.58%	73.77%	50.59%
0.4 Fine-tuned model on COLA + 0.6 Fine-tuned model on MRPC	0.4	3.64%	71.57%	51.35%
0.5 Fine-tuned model on COLA + 0.5 Fine-tuned model on MRPC	0.5	17.93%	48.04%	45.87%
0.6 Fine-tuned model on COLA + 0.4 Fine-tuned model on MRPC	0.6	46.69%	4.66%	37.56%
0.7 Fine-tuned model on COLA + 0.3 Fine-tuned model on MRPC	0.7	61.84%	0.00%	45.55%
0.8 Fine-tuned model on COLA + 0.2 Fine-tuned model on MRPC	0.8	65.58%	0.00%	48.31%
0.9 Fine-tuned model on COLA + 0.1 Fine-tuned model on MRPC	0.9	67.02%	0.00%	49.37%
1.0 Fine-tuned model on COLA + 0.0 Fine-tuned model on MRPC	1.0	67.88%	0.00%	50%

Figure 21: Finetune Middle Layer: COLA + MRPC

Model	Alpha	COLA	MRPC	Merging Performance
0.0 Fine-tuned model on COLA + 1.0 Fine-tuned model on MRPC	0	0.00%	73.53%	50%
0.1 Fine-tuned model on COLA + 0.9 Fine-tuned model on MRPC	0.1	1.05%	73.28%	50.87%
0.2 Fine-tuned model on COLA + 0.8 Fine-tuned model on MRPC	0.2	7.09%	70.83%	55.21%
0.3 Fine-tuned model on COLA + 0.7 Fine-tuned model on MRPC	0.3	16.78%	67.40%	62.50%
0.4 Fine-tuned model on COLA + 0.6 Fine-tuned model on MRPC	0.4	25.02%	61.27%	66.51%
0.5 Fine-tuned model on COLA + 0.5 Fine-tuned model on MRPC	0.5	30.11%	47.79%	62.40%
0.6 Fine-tuned model on COLA + 0.4 Fine-tuned model on MRPC	0.6	33.08%	19.12%	45.86%
0.7 Fine-tuned model on COLA + 0.3 Fine-tuned model on MRPC	0.7	36.63%	0.25%	36.55%
0.8 Fine-tuned model on COLA + 0.2 Fine-tuned model on MRPC	0.8	39.31%	0.00%	39.14%
0.9 Fine-tuned model on COLA + 0.1 Fine-tuned model on MRPC	0.9	44.30%	0.00%	44.00%
1.0 Fine-tuned model on COLA + 0.0 Fine-tuned model on MRPC	1.0	50.34%	0.00%	50%

Figure 22: Finetune Last Layer: COLA + MRPC