Supporting Document for jGSc

August 19, 2024

1 Discussion on Catastrophic Forgetting

Thank you for drawing attention to the work by Gueta et al. The results, references, and papers citing it show that our work fits within a broader line of work across modalities for model ensembling, and actually highlights one area where our work differs from the prior work we saw. We have added the following supplementary discussion, and the main text will refer to it when we discuss model ensembling:

“““ In our Results, we showed that CLIP models fine-tuned on pathology data (PLIP and QuiltNet) underperform their base models; we further showed that by interpolating weights between the fine-tuned and base models leads to a stronger general model, replicating a result by Wortsman et al 2022a with natural images. There is a large body of work on ensembling fine-tuned models with a base model, which we now discuss.

One difference of our result compared to from Wortsman et al 2022a (and most other works we saw) is that their fine-tuning phase has a different training strategy to pretraining: they do image classification trained with softmax loss. In our setting, the fine-tuning uses image-text pairs with the CLIP loss (the same as pretraining). This could be called ‘continued pretraining’ (Ibrahim et al 2024) on specialist data, but we chose to use ‘fine-tuned’ to be consistent with the past biomedical works, specifically the PLIP paper.

Gueta et al 2023 and Rame et al 2023 working with language models, and Wortsman et al 2022b working with CLIP, shows that superior performance can be achieved by ensembling multiple models fine-tuned on the same task with different data. This suggests that we could ensemble multiple CLIP models fine-tuned for pathology (PLIP and QuiltNet), however this is not possible because they use different base models with different architectures.

Another subtle difference is that the evaluation dataset we use for evaluation in MicroBench is using a different dataset to the fine-tuning dataset used in PLIP and QuiltNet, which is in line with some results by Chosen et al 2022, and Matena and Raffel 2021. ”””

As stated on our response, we would like to thank you again for bringing this relevant work to our attention. Given the challenges of curating high-quality, specialized microscopy data at the same scale as general Internet-based datasets (e.g., LAION 5B). Domain-specific biomedical VLMs often fine-tune a general-purpose model, which is a good practice. However, biomedical VLMs often overlook strategies to mitigate catastrophic forgetting, as highlighted by the fact that the specialist biomedical VLMs we evaluated did not initially employ such strategies. Overall, this expanded discussion emphasizes the importance of adopting strategies to mitigate catastrophic forgetting when fine-tuning VLMs for specialized biomedical applications.

References

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

[2] Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks. *arXiv preprint arXiv:2110.06296*, 2021.

[3] Almog Gueta, Elad Venezian, Colin Raffel, Noam Slonim, Yoav Katz, and Leshem Choshen. Knowledge is a region in weight space for fine-tuned language models. *arXiv preprint arXiv:2302.04863*, 2023.

[4] Adam Ibrahim, Benjamin Thérien, Kshitij Gupta, Mats L Richter, Quentin Anthony, Timothée Lesort, Eugene Belilovsky, and Irina Rish. Simple and scalable strategies to continually pre-train large language models. *arXiv preprint arXiv:2403.08763*, 2024.

[5] Michael S Matena and Colin A Raffel. Merging models with fisher-weighted averaging. *Advances in Neural Information Processing Systems*, 35:17703–17716, 2022.

[6] Yujia Qin, Cheng Qian, Jing Yi, Weize Chen, Yankai Lin, Xu Han, Zhiyuan Liu, Maosong Sun, and Jie Zhou. Exploring mode connectivity for pre-trained language models. *arXiv preprint arXiv:2210.14102*, 2022.

[7] Alexandre Ramé, Kartik Ahuja, Jianyu Zhang, Matthieu Cord, Léon Bottou, and David Lopez-Paz. Model ratatouille: Recycling diverse models for out-of-distribution generalization. In *International Conference on Machine Learning*, pages 28656–28679. PMLR, 2023.

[8] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pages 23965–23998. PMLR, 2022.

[9] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7959–7971, 2022.