Model Merging by Gradient Matching

Nico Daheim[1], **Thomas Möllenhoff**[2],
Edoardo M. Ponti[3], **Iryna Gurevych**[1], **Mohammad Emtiyaz Khan**[2]
[1]Ubiquitous Knowledge Processing Lab (UKP Lab)
Department of Computer Science and Hessian Center for AI (hessian.AI)
Technical University of Darmstadt
[2]RIKEN Center for Advanced Intelligence Project, Tokyo, Japan
[3]University of Edinburgh
`www.ukp.tu-darmstadt.de`

Abstract

Models trained on different datasets can be merged by a weighted-averaging of their parameters, but why does it work and when can it fail? Here, we connect the inaccuracy of weighted-averaging to mismatches in the gradients and propose a new uncertainty-based scheme to improve the performance by reducing the mismatch. The connection also reveals implicit assumptions in other schemes such as averaging, task arithmetic, and Fisher-weighted averaging.

1 Introduction

Merging models through a weighted averaging of their parameters has recently found many applications in deep learning. For example, averaging checkpoints generated during various training runs can improve out-of-distribution generalization (Izmailov et al., 2018; Wortsman et al., 2022b, *inter alia*), while averaging models trained on different datasets can borrow knowledge from "donor tasks" (Matena & Raffel, 2022) and enforce specific fine-grained behaviors in models (Ilharco et al., 2023).

The reasons behind the effectiveness of these methods are not well understood, and many schemes have been proposed, including arithmetic mean (Wortsman et al., 2022b,a), linear interpolation (Ilharco et al., 2023; Ortiz-Jimenez et al., 2023; Yadav et al., 2023), or individual parameter weighing (Matena & Raffel, 2022; Daheim et al., 2023). A prominent hypothesis, 'linear mode connectivity', is that when the models land in relatively few low-loss basins their interpolation again lies in them (Frankle et al., 2020; Neyshabur et al., 2020; Wortsman et al., 2022a; Ainsworth et al., 2023), but it does not tell us why one merging scheme should be preferred over the others or how to improve them.

In this abstract, we make two contributions: we first connect the inaccuracy of weighted-averaging to mismatches in the gradients and then improve its performance by reducing the mismatch with a second-order approximation; see an illustration in Fig. 1.

2 Model Merging by Parameter Averaging

We consider merging $T > 1$ models $\boldsymbol{\theta}_t \in \mathbb{R}^d$ with the same architecture that are trained on different datasets, for example, by fine-tuning a large pretrained model, such as $\boldsymbol{\theta}_{\text{LLM}}$. We focus on the following weighted-averaging scheme: $\bar{\boldsymbol{\theta}} = \mathbf{S}_0\, \boldsymbol{\theta}_{\text{LLM}} + \sum_{t=1}^T \mathbf{S}_t\, \boldsymbol{\theta}_t$, with scaling matrices $\mathbf{S}_t \in \mathbb{R}^{d\times d}$. Since d is often large, simple choices of $\mathbf{S}_t$ are used in practice, for example, scalars $\alpha_t > 0$ (Wortsman et al., 2022b,a), often tuned on held-out data. For large models, different parameters have different scaling and it is better to take this into account, for example, by using the Fisher $\mathbf{F}_t$ in 'Fisher Averaging': $\bar{\boldsymbol{\theta}}_{\text{FA}} = \sum_{t=1}^T \mathbf{S}_t\boldsymbol{\theta}_t,$ where $\mathbf{S}_t = \alpha_t\bar{\mathbf{F}}^{-1}\mathbf{F}_t$ with $\bar{\mathbf{F}} = \sum_{t=1}^T \alpha_t\mathbf{F}_t,$ for all $t \geq 1$ In practice, to reduce the computation cost, we may only use the diagonal of the Fisher estimated in

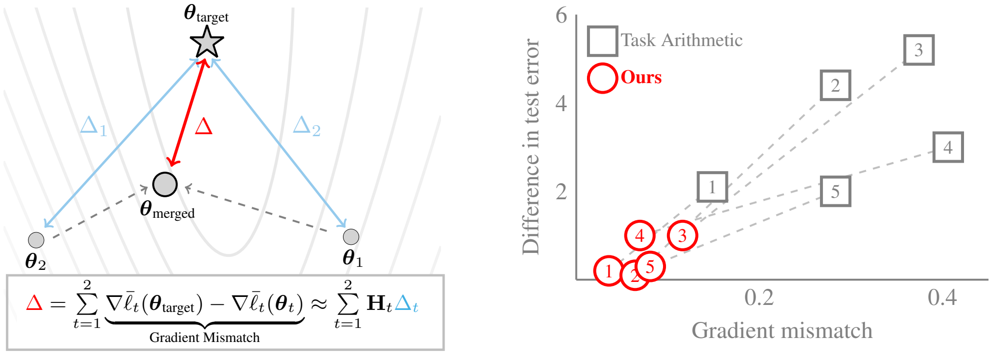

Figure 1: The left panel illustrates our approach. We connect the error Δ of the merged model $\boldsymbol{\theta}_{\text{merged}}$ to the gradient mismatch over losses $\bar{\ell}_t$ and propose a new method that reduces the mismatch by using the Hessian $\mathbf{H}_t$ and error Δ_t of the individual models $\boldsymbol{\theta}_t$. The right panel shows an example of adding datasets to RoBERTa trained on IMDB. We clearly see that reducing mismatch also reduces test error of task arithmetic. We consider 5 datasets, each indicated by a number on the markers.

an online fashion (Matena & Raffel, 2022). However, it is unclear how FA takes care of the commonalities of the $\mathbf{F}_t$ and $\boldsymbol{\theta}_{\text{LLM}}$. A recent work by Jin et al. (2023) uses insights from linear models to justify some choices, but may not hold for nonlinear models. Ilharco et al. (2023) proposed to subtract $\boldsymbol{\theta}_{\text{LLM}}$ with 'task arithmetic': $\bar{\boldsymbol{\theta}}_{\text{TA}} = \boldsymbol{\theta}_{\text{LLM}} + \sum_{t=1}^{T} \alpha_t(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{\text{LLM}})$, which reduces double-counting the information by using $\boldsymbol{\theta}_t - \boldsymbol{\theta}_{\text{LLM}}$, but it is unclear how to combine it with Fisher-style scaling.

We investigate the following: (1) how to choose scaling matrices; (2) what is the effect of these choices on the merged models' accuracy; and (3) how to obtain a new method that reduces inaccuracies.

3 Model Merging and Connections to Gradient Mismatches

To understand the inaccuracies of parameter averaging, we introduce the idea of a *target model*: it is the model that model merging methods want to estimate. Consider two models $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ trained on two datasets $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively, for example, as follows,

$$\boldsymbol{\theta}_1 = \arg\min_{\boldsymbol{\theta}} \; \bar{\ell}_1(\boldsymbol{\theta}) + \tfrac{1}{2}\|\boldsymbol{\theta}\|^2, \qquad \boldsymbol{\theta}_2 = \arg\min_{\boldsymbol{\theta}} \; \bar{\ell}_2(\boldsymbol{\theta}) + \tfrac{1}{2}\|\boldsymbol{\theta}\|^2. \tag{1}$$

Here, the loss functions on $\mathcal{D}_1$ and $\mathcal{D}_2$ are denoted by $\bar{\ell}_1(\boldsymbol{\theta})$ and $\bar{\ell}_2(\boldsymbol{\theta})$ respectively and the regularizer is an L_2 regularizer (what follows also holds for other explicit regularizers, also implicit ones). The target model in this case could be a model $\boldsymbol{\theta}_{1+2}$ that is trained jointly on the two datasets:

$$\boldsymbol{\theta}_{1+2} = \arg\min_{\boldsymbol{\theta}} \; \alpha_1\bar{\ell}_1(\boldsymbol{\theta}) + \alpha_2\bar{\ell}_2(\boldsymbol{\theta}_1) + \tfrac{1}{2}\|\boldsymbol{\theta}\|^2. \tag{2}$$

We will now connect gradient mismatch to the error between the target $\boldsymbol{\theta}_{1+2}$ and a parameter-average $\alpha_1\boldsymbol{\theta}_1 + \alpha_2\boldsymbol{\theta}_2$, but the approach is general and applies to different types of targets and averages.

We start with the first-order stationarity conditions of the models in Eqs. 1 and 2,

$$\boldsymbol{\theta}_1 = -\nabla\bar{\ell}_1(\boldsymbol{\theta}_1), \qquad \boldsymbol{\theta}_2 = -\nabla\bar{\ell}_2(\boldsymbol{\theta}_2), \qquad \boldsymbol{\theta}_{1+2} = -\alpha_1\nabla\bar{\ell}_1(\boldsymbol{\theta}_{1+2}) - \alpha_2\nabla\bar{\ell}_2(\boldsymbol{\theta}_{1+2}). \tag{3}$$

Using these, we can express $\boldsymbol{\theta}_{1+2}$ in terms of $\alpha_1\boldsymbol{\theta}_1 + \alpha_2\boldsymbol{\theta}_2$ and quantify the error made. To do so, we multiply the first and second equations above by α_1 and α_2 respectively, and add them together. Then, we subtract the resultant from the third equation to get the following expression:

$$\underbrace{\boldsymbol{\theta}_{1+2} - (\alpha_1\boldsymbol{\theta}_1 + \alpha_2\boldsymbol{\theta}_2)}_{=\Delta\text{, Error of the merged model}} = -\alpha_1 \underbrace{\left[\nabla\bar{\ell}_1(\boldsymbol{\theta}_{1+2}) - \nabla\bar{\ell}_1(\boldsymbol{\theta}_1)\right]}_{\text{Gradient mismatch for } \boldsymbol{\theta}_1 \text{ on } \bar{\ell}_1} - \alpha_2 \underbrace{\left[\nabla\bar{\ell}_2(\boldsymbol{\theta}_{1+2}) - \nabla\bar{\ell}_2(\boldsymbol{\theta}_2)\right]}_{\text{Gradient mismatch for } \boldsymbol{\theta}_2 \text{ on } \bar{\ell}_2}. \tag{4}$$

The left-hand side is the error $\Delta = \boldsymbol{\theta}_{1+2} - (\alpha_1\boldsymbol{\theta}_1 + \alpha_2\boldsymbol{\theta}_2)$ which is equal to the weighted-sum of the two gradient-mismatch terms on the individual losses $\bar{\ell}_1(\boldsymbol{\theta}_1)$ and $\bar{\ell}_2(\boldsymbol{\theta}_2)$. It shows that if the individual models are already close to the target model, parameter averaging should be reasonably accurate. It also shows us room for improvement and mismatch reduction may lead to better schemes.

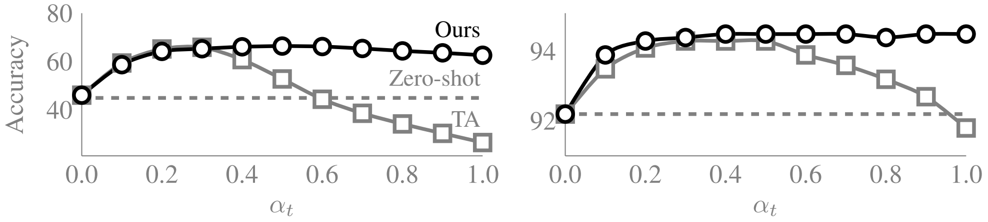

Figure 2: Our method is more robust to scaling than TA for task addition in CV (left) and NLP (right).

The method is generic and can be used to analyze errors of generic parameter-averaging schemes, for example, data removal (cf. Appendix). Test accuracy can also be analyzed. For example, given test loss $\bar{\ell}_{\text{Test}}(\boldsymbol{\theta})$ and weighted-average $\bar{\boldsymbol{\theta}}$, we have: $\bar{\ell}_{\text{Test}}(\boldsymbol{\theta}_{1+2}) - \bar{\ell}_{\text{Test}}(\bar{\boldsymbol{\theta}}) \approx \nabla\bar{\ell}_{\text{Test}}(\bar{\boldsymbol{\theta}})^\top(\boldsymbol{\theta}_{1+2} - \bar{\boldsymbol{\theta}})$. Large gradient mismatch therefore is expected to correlate with large differences in test performance.

Sources of errors can be analyzed, too. For example, when the test data is more correlated to $\mathcal{D}_1$, then model merging would be effective if gradient mismatch due to $\boldsymbol{\theta}_1$ is also small. This is similar to linear mode connectivity: when both the target and merged models lie in low-loss basins, we expect gradient mismatch to be low due to flatness. However, gradient-mismatch does not require this and is more general and constructive by allowing us to improve models by actively reducing the mismatch.

3.1 Analyzing the Inaccuracy of Task Arithmetic on Large Language Models

We will demonstrate the use of the gradient-mismatch principle to analyze the inaccuracy of ‘task arithmetic’ (Ilharco et al., 2023) for $\boldsymbol{\theta}_{\text{LLM}}$ trained on a large dataset $\mathcal{D}_{\text{Large}}$.

$$\boldsymbol{\theta}_{\text{LLM}} = \arg\min_{\boldsymbol{\theta}} \ \bar{\ell}_{\text{LLM}}(\boldsymbol{\theta}) + \tfrac{1}{2}\delta\|\boldsymbol{\theta}\|^2, \text{ where } \bar{\ell}_{\text{LLM}}(\boldsymbol{\theta}) = \sum_{i\in\mathcal{D}_{\text{Large}}} \ell_i(\boldsymbol{\theta}). \quad (5)$$

Here, $\ell_i(\boldsymbol{\theta})$ denotes the loss on the i’th example. For simplicity, we use an L_2 regularization with parameter $\delta > 0$ but the choice is not crucial. The loss function can also be normalized. Our goal is to merge models $\boldsymbol{\theta}_t$ that are finetuned on different datasets $\mathcal{D}_t$ for $t = 1, 2, \ldots, T$ using:

$$\boldsymbol{\theta}_t = \arg\min_{\boldsymbol{\theta}} \ \bar{\ell}_t(\boldsymbol{\theta}) + \tfrac{1}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{LLM}}\|^2_{\mathbf{H}_0}, \quad (6)$$

where $\|\boldsymbol{\theta}\|^2_{\mathbf{H}_0} = \boldsymbol{\theta}^\top\mathbf{H}_0\boldsymbol{\theta}$ is the Mahalanobis distance with a scaling matrix $\mathbf{H}_0$ which controls how different $\boldsymbol{\theta}$ is from $\boldsymbol{\theta}_{\text{LLM}}$. We will discuss how to set $\mathbf{H}_0$ later. *The derivation can be easily adopted to other fine-tuning procedures* as long as we can express the dependence on $\boldsymbol{\theta}_{\text{LLM}}$ explicitly.

Task arithmetic (TA) uses $\bar{\boldsymbol{\theta}}_{\text{TA}} = \boldsymbol{\theta}_{\text{LLM}} + \sum_t \alpha_t(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{\text{LLM}})$ to merge models. There are two natural questions: what is the target model that such a scheme is trying to approximate and what are the errors made by TA in approximating it? As before, a reasonable choice of the target model is the one obtained by fine-tuning using a similar procedure as Eq. 6 but on all $\mathcal{D}_t$ at once,

$$\boldsymbol{\theta}_{1:T} = \arg\min_{\boldsymbol{\theta}} \ \sum_{t=1}^{T} \alpha_t\ell_t(\boldsymbol{\theta}) + \tfrac{1}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{LLM}}\|^2_{\mathbf{H}_0}. \quad (7)$$

Following the same derivation as Eq. 4, we can quantify the error between $\boldsymbol{\theta}_{1:T}$ and $\bar{\boldsymbol{\theta}}_{\text{TA}}$ (a full derivation is given in Appendix):

$$\boldsymbol{\theta}_{1:T} = \underbrace{\boldsymbol{\theta}_{\text{LLM}} + \sum_{t=1}^{T}\alpha_t(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{\text{LLM}})}_{=\bar{\boldsymbol{\theta}}_{\text{TA}}} - \sum_{t=1}^{T}\alpha_t\mathbf{H}_0^{-1}\underbrace{\left[\nabla\bar{\ell}_t(\boldsymbol{\theta}_{1:T}) - \nabla\bar{\ell}_t(\boldsymbol{\theta}_t)\right]}_{\text{Gradient mismatch for } \boldsymbol{\theta}_t \text{ on } \bar{\ell}_t}. \quad (8)$$

The derivation can be used to understand the implicit assumptions made in task arithmetic. The increments $\boldsymbol{\theta}_t - \boldsymbol{\theta}_{\text{LLM}}$ arise above due to the quadratic regularizer $\|\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{LLM}}\|^2$ used in Eqs. 6 and 7 and avoid double counting. More importantly, the error between $\boldsymbol{\theta}_{1:T}$ and $\bar{\boldsymbol{\theta}}_{\text{TA}}$ is attributed to gradient mismatch between $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}_{1:T}$. The expression suggests that by reducing the mismatch we could improve task arithmetic. We will now show that a simple method that uses Taylor’s approximation to reduce the gradient mismatch justifies combining TA with a Fisher-like weighting schemes.

	IMDB	Yelp	RT	SST2	Amazon	Avg.
All-data	94.8	97.6	91.2	94.7	96.9	95.0
Averaging	94.4	97.0	89.1	93.6	96.2	94.1
Fisher Averaging	**94.8**	97.2	89.9	93.1	96.6	94.3
Task Arithmetic (tuned α_t)†	94.3	97.2	89.6	**94.5**	96.4	94.4
Ours	94.7 (↑0.4)	**97.3** (↑0.1)	**90.2** (↑0.6)	93.7 (↓0.8)	**96.7** (↑0.3)	**94.5** (↑0.1)

Table 1: We merge four tasks with RoBERTa trained on IMDB. Our merging function shows that reducing gradient mismatch improves performance over previously proposed functions.

3.2 A New Method to Reduce the Gradient Mismatch

We now derive a new parameter-averaging method by reducing the gradient mismatch in Eq. 8. Explicit minimization of the mismatch is non-trivial because $\nabla\bar{\ell}_t(\boldsymbol{\theta}_{1:T})$ depends non-linearly on $\boldsymbol{\theta}_{1:T}$ but we can get rid of the term by using a first-order Taylor approximation,

$$\nabla\bar{\ell}_t(\boldsymbol{\theta}) \approx \nabla\bar{\ell}_t(\boldsymbol{\theta}_t) + \mathbf{H}_t(\boldsymbol{\theta} - \boldsymbol{\theta}_t) \tag{9}$$

where $\mathbf{H}_t = \nabla^2\bar{\ell}_t(\boldsymbol{\theta}_t)$ is the Hessian of the loss $\bar{\ell}_t$ at $\boldsymbol{\theta}_t$. Using this in Eq. 8 and after some rearrangement, we get the following merging scheme (a full derivation is given in Appendix),

$$\hat{\boldsymbol{\theta}}_{1:T} = \boldsymbol{\theta}_{\text{LLM}} + \sum_{t=1}^{T} \alpha_t \, (\bar{\mathbf{H}}^{-1}\mathbf{H}_{0+t}) \, (\boldsymbol{\theta}_t - \boldsymbol{\theta}_{\text{LLM}}), \tag{10}$$

where $\bar{\mathbf{H}} = \mathbf{H}_0 + \sum_{t=1}^{T} \alpha_t\mathbf{H}_t$ and $\mathbf{H}_{0+t} = \mathbf{H}_0 + \mathbf{H}_t$ is the Hessian plus a regularization matrix. The new merging scheme adds preconditioners $\bar{\mathbf{H}}^{-1}\mathbf{H}_{0+t}$ to task arithmetic. The preconditioners depend on the Hessians $\mathbf{H}_t$, which is similar to Fisher averaging, but here the choice naturally emerges as a consequence of the gradient-mismatch reduction. Nevertheless, replacing $\mathbf{H}_t$ by the diagonal Fisher $\mathbf{F}_t$ of $\boldsymbol{\theta}_t$ is often easier to compute and easier numerically because positive-definiteness is ensured. The matrix $\mathbf{H}_0$ can be set in a similar way, for example, to the Hessian/Fisher of Eq. 5 at $\boldsymbol{\theta}_{\text{LLM}}$.

Choosing different setting of α_t, $\mathbf{H}_0$, and $\mathbf{H}_t$, can recover many existing schemes as special cases of Eq. 10. This helps us to understand not only their inaccuracies but also their implicit assumptions. AM and TA can be seen as special cases where the preconditioner $\mathbf{H}_t = \mathbf{0}$. This implies that the gradient mismatch term in Eq. 8 is left as is and the error will be high when there gradient mismatch is high. In contrast, Fisher averaging can be seen as a special cases where $\mathbf{H}_0 = \mathbf{0}$ which implies that the quadratic regularizer in Eqs. 6 and 7 vanishes, ignoring the dependence of $\boldsymbol{\theta}_t$ on $\boldsymbol{\theta}_{\text{LLM}}$.

4 Experiments & Results

We use a pretrained ViT (Dosovitskiy et al., 2021) for image classification and add eight datasets to it: Cars (Krause et al., 2013), DTD (Cimpoi et al., 2014), EuroSAT (Helber et al., 2018), GTSRB (Houben et al., 2013), MNIST (LeCun, 1998), RESISC45 (Cheng et al., 2017), SUN397 (Xiao et al., 2010), and SVHN (Yuval, 2011), replicating Ilharco et al. (2023). We use identity to approximate the Hessian of the ViT and all task models are trained by fine-tuning it. The results are outlined in the leftmost panel of Fig. 2. Our proposed merging function is much more robust to the choice of scaling factors. For larger factors, task arithmetic even falls below the zero-shot baseline.

We repeat a similar experiment with RoBERTa (Liu et al., 2019) for sentiment classification, which we first train on IMDB (Maas et al., 2011) (arbitrarily chosen). We approximate $\mathbf{H}_0$ using squared gradients on the training data. We then use this model to initialize all $\boldsymbol{\theta}_t$ which we train on Amazon (Zhang et al., 2015), RottenTomatoes (Pang & Lee, 2005), SST2 (Socher et al., 2013), and Yelp (Zhang et al., 2015). Table 1 shows that our new method gets closer to the "all-data" target model than other merging functions, indicating that reducing gradient mismatch is crucial, as outlined also in Fig. 1. Furthermore, it improves over TA even when we tune scaling factors on the test set for TA and not at all for our method. Fig. 2 (right) shows a plot over scaling factors where our method dominates TA which also falls below the zero-shot baseline of the IMDB model.

5 Conclusion

We have connected the error of the merged model to the gradient mismatch between the individual models that are merged and the 'target model' that merging aims to recover. We have used this to reveal implicit assumptions in related methods and propose an improved merging scheme that is more robust in terms of scaling factors and improves downstream performance.

Acknowledgements

This project has received funding by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE. This work is supported by the Bayes duality project, JST CREST Grant Number JPMJCR2112.

References

Samuel Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git Re-Basin: Merging models modulo permutation symmetries. In *International Conference on Learning Representations (ICLR)*, 2023. URL `https://openreview.net/forum?id=CQsmMYmlP5T`. pages 1

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proceedings of the IEEE*, 105(10):1865–1883, Oct 2017. ISSN 1558-2256. doi: 10.1109/jproc.2017.2675998. URL `http://dx.doi.org/10.1109/JPROC.2017.2675998`. pages 4

M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi. Describing Textures in the Wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014. pages 4

Nico Daheim, Nouha Dziri, Mrinmaya Sachan, Iryna Gurevych, and Edoardo M. Ponti. Elastic weight removal for faithful and abstractive dialogue generation, 2023. URL `https://arxiv.org/abs/2303.17574`. arXiv:2303.17574. pages 1

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`. pages 4

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning (ICML)*, 2020. pages 1

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Introducing EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. In *IEEE International Geoscience and Remote Sensing Symposium*, 2018. pages 4

Sebastian Houben, Johannes Stallkamp, Jan Salmen, Marc Schlipsing, and Christian Igel. Detection of traffic signs in real-world images: The German Traffic Sign Detection Benchmark. In *International Joint Conference on Neural Networks (IJCNN)*, 2013. pages 4

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations (ICLR)*, 2023. URL `https://openreview.net/forum?id=6t0Kwf8-jrj`. pages 1, 2, 3, 4

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2018. URL `http://auai.org/uai2018/proceedings/papers/313.pdf`. pages 1

Xisen Jin, Xiang Ren, Daniel Preotiuc-Pietro, and Pengxiang Cheng. Dataless knowledge fusion by merging weights of language models. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=FCnohuR6AnM. pages 2

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *International Conference on Computer Vision (ICCV)*, 2013. (Workshops). pages 4

Yann LeCun. The MNIST database of handwritten digits, 1998. URL http://yann.lecun.com/exdb/mnist/. pages 4

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach, 2019. URL http://arxiv.org/abs/1907.11692. arXiv:1907.11692. pages 4

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2011. URL http://www.aclweb.org/anthology/P11-1015. pages 4

Michael S Matena and Colin Raffel. Merging models with Fisher-weighted averaging. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://openreview.net/forum?id=LSKlp_aceOC. pages 1, 2

Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. pages 1

Guillermo Ortiz-Jimenez, Alessandro Favero, and Pascal Frossard. Task arithmetic in the tangent space: Improved editing of pre-trained models, 2023. URL http://arxiv.org/abs/2305.12827. arXiv:2305.12827. pages 1

Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2005. pages 4

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2013. URL https://www.aclweb.org/anthology/D13-1170. pages 4

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning (ICML)*, 2022a. URL https://proceedings.mlr.press/v162/wortsman22a.html. pages 1

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022b. https://arxiv.org/abs/2109.01903. pages 1

J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010. pages 4

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. Resolving interference when merging models, 2023. URL http://arxiv.org/abs/2306.01708. arXiv:2306.01708. pages 1

Netzer Yuval. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011. pages 4

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level Convolutional Networks for Text Classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015. pages 4