
ZipIt!: Multitask Model Merging *without Training*

Anonymous Author(s)

Affiliation
Address
email

Abstract

1 We tackle the extremely difficult problem of combining distinct models with
2 different initializations, each solving a separate task, into one multi-task model
3 **without any additional training**. Prior work in model merging permutes one
4 model to the space of the other then averages them together. While this works for
5 models trained on the same task, we find that this fails to account for the differences
6 in models trained on disjoint tasks. Thus, we introduce “ZipIt!”, a general method
7 for merging two arbitrary models of the same architecture that incorporates two
8 simple strategies. First, in order to account for features that aren’t shared between
9 models, we expand the model merging problem to allow for merging features
10 *within* each model by defining a general “zip” operation. Second, we add support
11 for *partially zipping* the models up until a specified layer, naturally creating a multi-
12 head model. We find that these two changes combined account for a staggering
13 20-50% improvement over prior work,

14

1 Introduction

15 Combining multiple models into one *without training* has recently started to gain traction in the
16 vision community. Model Soups [1] can add multiple models finetuned from the same pretrained
17 initialization to improve accuracy and robustness. Git Re-Basin [2] generalizes further to models
18 trained on the same data but with different initializations, though with a significant accuracy drop.
19 REPAIR [3] improves on Git Re-Basin by adding new parameters and adjusting model batch norms
20 where applicable. However, all of these methods only combine models trained on the same task.
21 In this paper, we take this line of work to a logical extreme: merging differently initialized models
22 trained on *completely separate* tasks (see Fig. 1ab). We show that this is an incredibly difficult
23 problem for prior work and employ two simple strategies to make it feasible.

24 First, prior work focuses on *permuting* one model to the other when merging them. This inherently
25 assumes that most features *across* them are correlated. But, this is not always the case for models
26 trained on different tasks. Instead, we generalize model merging to support “zipping” any combination
27 of correlated features *within* and *across* each model. We find that on some tasks, this alone improves
28 accuracy **by up to 20%** vs. permutation-based approaches

29 Second, the features of models trained on disjoint tasks become less correlated over the course of the
30 network [4]. Thus, we introduce *partial zipping*, which allows us to only “zip” up to a specified layer.
31 Afterward, we feed the merged model’s outputs to the remaining unmerged layers of the original
32 networks, creating a multi-head model. Partially zipping can improve accuracy **by over 15%**.

33 ZipIt! (Fig. 1c) incorporates both strategies to “zip” models trained on different tasks into a single
34 multitask model *without retraining*. ZipIt! is general and supports merging arbitrary models of the
35 same architecture together (Sec. 3). We validate ZipIt! by merging models trained on different tasks
36 (including classification and segmentation) into one, significantly outperforming prior work (Sec. 4).

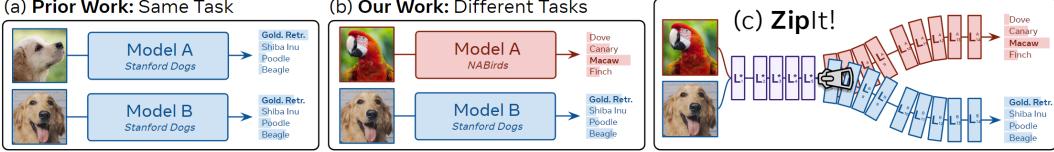


Figure 1: **Setting and ZipIt!** (a) Prior work merges models from the **same dataset** with the **same label sets**: e.g., merging two models both trained to classify dog breeds. (b) Our setting expands to merging models from **different datasets** with **different label sets**: e.g., merging a model that classifies dog breeds with one that classifies bird species. (c) **ZipIt!** merges these models *without retraining* by identifying shared features. Depending on the task, ZipIt! can nearly match ensemble performance.

37 2 Related Work

38 Model merging combines the weights of two or more models into a one. Our work differs from prior
39 work in that we merge differently initialized models trained on disjoint tasks (Fig. 1) without training.

40 **Merging Finetuned Models.** If two models are finetuned from the same pretrained checkpoint,
41 they often lie in the same error minima [5]. [6, 7, 8, 9, 10, 11] have exploited this property to
42 average together the weights of a model at different stages of training. [12, 13, 14, 15, 16] use an
43 “exponential moving average” of training checkpoints as a teacher for additional self-supervised
44 learning. Other works, [1, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27] merge models fully finetuned to
45 improve performance on a task. Our setting differs, as we do not assume the same initialization.

46 **Merging Differently Initialized Models.** Works in this space often merge models trained on the
47 same task and rely on mode connectivity [28, 29, 30, 31, 32, 33, 34], as differently initialized models
48 may not lie in the same error minima [29, 35]. Most recent work follows the intuition formalized
49 by [35] that models permuted to the same loss minima can be merged by averaging their weights
50 [35, 2, 3, 36]. Similar to ZipIt! [37] merges models of different tasks, but requires jointly finetuning
51 on all tasks after each layer merge. As far as we are aware, we present the first *general method* to
52 successfully merge models trained on disjoint tasks *without additional training*.

53 3 ZipIt!

54 In this work, we treat model merging as combining the checkpoints (i.e., collection of weights) of
55 multiple models layer-by-layer into a single checkpoint that can perform all the tasks of its constituents.
56 Consider a model \mathcal{L} as a collection of layers $L_i \in \mathcal{L}$, each of which may have some parameters (e.g.,
57 W_i, b_i for a linear layer). Suppose $L_i \in \mathcal{L}$ is a linear layer with parameters $W_i \in \mathbb{R}^{n_i \times m_i}, b_i \in \mathbb{R}^{n_i}$
58 with input features $x \in \mathbb{R}^{m_i}$ and outputs features $f_i \in \mathbb{R}^{n_i}$. Then, $f_i = L_i(x) = W_i x + b_i$.

59 Our goal is to take $L_i^A \in \mathcal{L}^A$ from a **model A** and $L_i^B \in \mathcal{L}^B$ from a **model B** and merge them into a
60 layer L_i^* that combines their feature spaces such that information from both f_i^A and f_i^B is retained
61 in f_i^* . We accomplish this by merging each layer of one model with the corresponding layer in the
62 other, both merging features *across* both layers and *within* the same layer. This is in contrast to
63 permutation-based merging method, which only combine features *across* layers.

64 **Problems with Permutation.** Permutation methods posit that **model B** can be moved to the same
65 loss minima as **model A** via permutation with high-likelihood [35]. However, this is unlikely when
66 models are trained on different tasks because each model optimizes for task-specific minimas. In this
67 case the optimal permutation of **model B** to **model A** lies in a strong minima on **task B** but *may not*
68 lie in a minima on **task A**, as shown in Fig. 2. This causes the interpolated model to perform worse
69 than either of the two original models. Thus, we explore alternative merging methods.

70 **Why should we merge *within*?** Features of models trained on different tasks may be dissimilar, as
71 the models solve different problems. Instead, those features may be more compatible with others
72 within the same model, which would better retain performance when combined.

73 **What is our merge strategy?** Prior work obtains f_i^* by combining one feature from f_i^A and one
74 from f_i^B . [38, 3] determine which features to merge by computing the pairwise correlations between



Figure 2: **Task Loss Landscapes** for models in Tab. 1b. Model A and Model B lie in low loss minimas for their own tasks, but *not for the other task*. Thus, any interpolation between Model A and a permuted Model B (e.g., Git Re-basin) lies outside the minima *for both tasks* and thus performs poorly. In contrast, ZipIt! improves the merge by finding a model that lies in a loss minima for both.

75 the neuron activations of f_i^A and f_i^B over a set of images. In contrast, our approach can also obtain
 76 f_i^* by combining two features from just f_i^A or f_i^B . We compute the same correlations, but also
 77 include those of f_i^A and f_i^B each with themselves over the same data. We then greedily choose
 78 without replacement the pairs with highest correlation and average their features to obtain f_i^* .

79 **Merging a layer.** For every layer, we first concatenate f_i^A and f_i^B into $f_i^A \| f_i^B \in \mathbb{R}^{2n_i}$, and
 80 define a ‘‘merge matrix’’ $M_i \in \mathbb{R}^{n_i \times 2n_i}$ based on the matches from our greedy algorithm (i.e.,
 81 $f_i^* = M_i(f_i^A \| f_i^B)$). Thus, M_i merges the output spaces of L_i^A , and L_i^B into L_i^* . We then define an
 82 ‘‘unmerge’’ matrix $U_i \in \mathbb{R}^{2n_i \times n_i}$ s.t. it *undoes* the merge operation: $U_i f_i^* \approx f_i^A \| f_i^B$. Applying U_i
 83 ensures the input spaces of the future layers L_{i+1}^A , and L_{i+1}^B are aligned with current merged layer
 84 L_i^* . Once all matrices are computed at each layer, we obtain L_i^* for a linear layer by

$$W_i^* = M_i \begin{pmatrix} W_i^A & 0 \\ 0 & W_i^B \end{pmatrix} U_{i-1} \quad b_i^* = M_i \begin{pmatrix} b_i^A \\ b_i^B \end{pmatrix}.$$

85 Due to space restrictions, please see Appendix. A for more details on merging other kinds of layers.

86 **Partial Zipping.** Sometimes, later layers in the networks have very uncorrelated (dissimilar) outputs.
 87 Forcibly zipping these would lead to meaningless features. In this case, we can perform a *partial zip*.
 88 That is, we zip most of the layers together, but leave the later ones *unzipped*—obtaining a multihead
 89 model.

90 **Merging Many Models (α).** Sometimes, we’d like to merge more than two models together. To
 91 do this, we allow ‘‘repeated matches’’: we replace matched features from our algorithm with the
 92 resulting merged feature instead of removing them completely. To ensure that one feature doesn’t get
 93 merged endlessly, we set the correlations of the new feature to be the minimum of the old features’
 94 similarities weighted by $\alpha \in (0, 1]$. We find a small value of α typically works best.

95 **Within-Model Merging Budget (β).** To demonstrate the effectiveness of same-model merges, we
 96 introduce a ‘‘budget’’ parameter $\beta \in [0, 1]$ that denotes what percent of total merged features can
 97 come from models merging within themselves, with each model receiving an equal portion of this
 98 budget. A budget of 0 only allows feature merging across models and yields a permutation-merge.

99 4 Results

100 There’s no standard benchmark to evaluate merging approaches on models from distinct tasks, so
 101 we construct our own. We evaluate our approach in two different settings. (1) A versatile test-bed:
 102 disjoint category splits of the same dataset (i.e., *same dataset and different label sets*). (2) A very
 103 challenging setting: completely different datasets and tasks (i.e., *different datasets and label sets*).

104 We compare to three baselines: Git Re-Basin [2], Weight Averaging (W. Avg) and Permute. W. Avg
 105 involves simply average all model parameters to be merged, while we design Permute to use linear
 106 sum assignment to find optimal permutations (following [38]) for merging. Note that Permute is a

Method	FLOPs (G)	Accuracies (%)			
		Joint	Task A	Task B	Avg
Model A	0.68	48.2 \pm 0.1	97.0 \pm 0.6	45.1 \pm 8.6	71.0 \pm 4.4
Model B	0.68	48.4 \pm 3.8	49.1 \pm 0.3	96.1 \pm 1.1	72.6 \pm 4.9
W. Avg	0.68	43.0 \pm 1.6	54.1 \pm 1.4	67.5 \pm 1.2	60.8 \pm 4.5
Git Re-Basin [†]	0.68	46.2 \pm 0.8	76.8 \pm 8.9	82.7 \pm 5.1	79.8 \pm 6.5
Permute	0.68	58.4 \pm 6.8	86.6 \pm 2.1	87.4 \pm 1.1	87.4 \pm 1.4
ZipIt!_{20/20}	0.68	79.1 \pm 1.1	92.9 \pm 1.1	91.2 \pm 1.4	92.1 \pm 1.0
Ensemble	1.37	87.4 \pm 2.6	97.0 \pm 0.6	96.1 \pm 1.1	96.6 \pm 0.4
ZipIt!_{13/20}	0.91	83.8\pm3.1	95.1\pm0.7	94.1\pm1.5	94.6\pm0.6

(a) CIFAR-10 (5+5). ResNet-20 ($4 \times$ width).

Method	FLOPs (G)	Accuracies (%)			
		Joint	Task A	Task B	Avg
Model A	2.72	41.6 \pm 0.3	82.9 \pm 0.7	24.8 \pm 0.4	53.9 \pm 0.5
Model B	2.72	41.6 \pm 0.2	25.1 \pm 1.2	82.8 \pm 0.2	54.0 \pm 0.6
W. Avg	2.72	17.0 \pm 1.7	23.8 \pm 6.9	24.8 \pm 5.9	24.3 \pm 1.9
Git Re-Basin [†]	2.72	40.9 \pm 0.2	57.3 \pm 1.5	56.7 \pm 0.7	57.0 \pm 0.8
Permute	2.72	42.8 \pm 0.7	61.6 \pm 1.4	60.5 \pm 0.5	61.0 \pm 0.8
ZipIt!_{20/20}	2.72	54.9\pm0.8	68.2\pm0.8	67.9\pm0.6	68.0\pm0.4
Ensemble	5.45	73.5 \pm 0.4	82.9 \pm 0.7	82.8 \pm 0.2	82.8 \pm 0.4
ZipIt!_{13/20}	3.63	70.2\pm0.4	80.3\pm0.8	80.1\pm0.7	80.2\pm0.6

(b) CIFAR-100 (50+50). ResNet-20 ($8 \times$ width).

Table 1: **CIFAR Results.** ZipIt! vs. baselines on combining a model trained on half the classes (**Task A**) with one trained on the other half (**Task B**) *without extra training*. We report both joint (10/100-way) and per-task (5/50-way) accuracy. ZipIt! significantly outperforms its baseline and closes in on the upper bound (ensemble accuracy). [†] refers to [2].

107 *strong* baseline we create and is more accurate than Git Re-Basin in our settings. For our method,
108 ZipIt!_{n/m} indicates that n out of the m layers in the network have been zipped (Sec. 3). Note, all our
109 models have *different initializations*.

110 **Disjoint Category Splits.** In Tab. 1a, we merge five pairs of models trained on disjoint 5 class subsets
111 of CIFAR-10 using ResNet-20 with a $4 \times$ width multiplier (denoted as ResNet-20 \times 4). We train with
112 a CLIP-style loss [39] using CLIP text encodings of the class names as targets so that both models
113 output into the same CLIP-space regardless of the category (required for [2]). We report: (1) joint
114 accuracy—the accuracy of each model over *all* classes across datasets, and (2) per task accuracy—the
115 accuracy of each task individually and also their average. Overall, ZipIt! performs a staggering 20.7%
116 better than the nearest baseline. If we allow the last stage of the network to remain unzipped (i.e., zip
117 up to 13 layers), our method obtains 83.8%, which is only 3.6% behind an ensemble of **model A** and
118 **model B** (which is practically the upper bound for this setting). We find similar results on disjoint
119 50 class splits of CIFAR-100 in Tab. 1b using an $8 \times$ width multiplier. ZipIt! again significantly
120 outperforms baselines.

121 **Different Datasets.** We merge ResNet-50 mod-
122 els trained on: Stanford Dogs [40], Oxford Pets
123 [41], CUB200 [42], and NABirds [43]. In Tab. 2,
124 we show the average per task accuracy from ex-
125 haustively merging each pair and the much more
126 difficult setting of merging all four at once. We
127 report the accuracy of our baselines by applying
128 them up until the last layer, but we can't compare
129 to [2] as it requires shared output space.

130 For pairs of models, ZipIt! slightly outperforms
131 our permute baseline across all tasks and performs
132 similarly when merging all 4 models at once. How-
133 ever, if we add capacity to the merged model
134 through partial zipping, we perform up to 33%
135 better on merging pairs and 50% better on merg-
136 ing all four models than the permute baseline.

137 We also merge the ResNet-50 backbone of a
138 DeeplabV3 [44] segmentation model that achieves
139 76.8% mIoU on PASCAL VOC [45] with an ImageNet-1k [46] classification model that obtains
140 77.8% accuracy. ZipIt! can still achieve good performance on *both* tasks even with half of layers
141 merged: obtaining 64.4% mIoU on PASCAL VOC and 60.9% accuracy on ImageNet-1k.

Method	FLOPs (G)	Per-Task Accuracies (%)				
		Merging Pairs	SD	OP	CUB	NAB
Merging All 4						
W. Avg	4.11	12.9	18.2	13.9	0.2	11.3
Permute	4.11	46.2	47.6	35.6	13.5	35.7
ZipIt!_{49/50}	4.11	46.9	50.7	38.0	12.7	37.1
Ensemble	8.22	72.7	81.1	71.0	77.2	75.5
ZipIt!_{22/50}	6.39	62.6	71.2	62.8	53.0	62.4
ZipIt!_{10/50}	7.42	66.5	75.8	65.6	66.8	68.7

Table 2: **Multi-Dataset Results.** Merging ResNet-50 models trained on *completely different datasets*: Stanford Dogs (SD), Oxford Pets (OP), CUB200 (CUB), and NABirds (NAB). We report average per-task accuracy over merging model pairs, and all four.

142 5 Conclusion

143 In this paper, we tackle the extremely difficult task of merging models trained on completely disjoint
144 tasks *without additional training*. We find that prior work underperforms in this setting and posit
145 that they neither fully (1) exploit model similarities nor (2) account for model dissimilarities. We
146 introduce ZipIt!, a general framework for merging models that addresses these issues, and show it to
147 significantly outperform prior work across several difficult settings.

148 **References**

- 149 [1] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes,
150 Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model
151 soups: averaging weights of multiple fine-tuned models improves accuracy without increasing
152 inference time. In *ICML*, 2022.
- 153 [2] Samuel K Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models
154 modulo permutation symmetries. *arXiv:2209.04836*, 2022.
- 155 [3] Keller Jordan, Hanie Sedghi, Olga Saukh, Rahim Entezari, and Behnam Neyshabur. Repair:
156 Renormalizing permuted activations for interpolation repair. *arXiv:2211.08403*, 2022.
- 157 [4] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural
158 network representations revisited. In *ICML*, 2019.
- 159 [5] Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer
160 learning? *NeurIPS*, 2020.
- 161 [6] Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E Hopcroft, and Kilian Q Weinberger.
162 Snapshot ensembles: Train 1, get m for free. *arXiv:1704.00109*, 2017.
- 163 [7] Pavel Izmailov, Dmitrii Podoprikin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon
164 Wilson. Averaging weights leads to wider optima and better generalization. *UAI*, 2018.
- 165 [8] Johannes Von Oswald, Seijin Kobayashi, Joao Sacramento, Alexander Meulemans, Christian
166 Henning, and Benjamin F Grewe. Neural networks with late-phase weights. *arXiv:2007.12927*,
167 2020.
- 168 [9] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca
169 Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al.
170 Robust fine-tuning of zero-shot models. In *CVPR*, 2022.
- 171 [10] Gabriel Ilharco, Mitchell Wortsman, Samir Yitzhak Gadre, Shuran Song, Hannaneh Hajishirzi,
172 Simon Kornblith, Ali Farhadi, and Ludwig Schmidt. Patching open-vocabulary models by
173 interpolating weights. *NeurIPS*, 2022.
- 174 [11] Shachar Don-Yehiya, Elad Venezian, Colin Raffel, Noam Slonim, and Leshem Choshen. Cold
175 fusion: Collaborative descent for distributed multitask finetuning. Association for Computational
176 Linguistics, 2023.
- 177 [12] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged
178 consistency targets improve semi-supervised deep learning results. *NeurIPS*, 2017.
- 179 [13] Zhaowei Cai, Avinash Ravichandran, Subhransu Maji, Charless Fowlkes, Zhuowen Tu, and
180 Stefano Soatto. Exponential moving average normalization for self-supervised and semi-
181 supervised learning. In *CVPR*, 2021.
- 182 [14] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena
183 Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar,
184 et al. Bootstrap your own latent-a new approach to self-supervised learning. *NeurIPS*, 2020.
- 185 [15] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski,
186 and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.
- 187 [16] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli.
188 Data2vec: A general framework for self-supervised learning in speech, vision and language. In
189 *ICML*, 2022.
- 190 [17] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas.
191 Communication-efficient learning of deep networks from decentralized data. In *Artificial
192 intelligence and statistics*. PMLR, 2017.
- 193 [18] Leshem Choshen, Elad Venezian, Noam Slonim, and Yoav Katz. Fusing finetuned models for
194 better pretraining. *arXiv:2204.03044*, 2022.

- 195 [19] Alexandre Ramé, Kartik Ahuja, Jianyu Zhang, Matthieu Cord, Léon Bottou, and David Lopez-
196 Paz. Recycling diverse models for out-of-distribution generalization. *arXiv:2212.10445*, 2022.
- 197 [20] Stephen Ashmore and Michael Gashler. A method for finding similarity between multi-layer
198 perceptrons by forward bipartite alignment. In *IJCNN*, 2015.
- 199 [21] Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Nghia Hoang, and
200 Yasaman Khazaeni. Bayesian nonparametric federated learning of neural networks. In *ICML*,
201 2019.
- 202 [22] Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris Papailiopoulos, and Yasaman Khazaeni.
203 Federated learning with matched averaging. *arXiv:2002.06440*, 2020.
- 204 [23] Gabriel Ilharco, Marco Túlio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt,
205 Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *arXiv:2212.04089*,
206 2022.
- 207 [24] Almog Gueta, Elad Venezian, Colin Raffel, Noam Slonim, Yoav Katz, and Leshem Choshen.
208 Knowledge is a region in weight space for fine-tuned language models. *arXiv:2302.04863*,
209 2023.
- 210 [25] Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. Resolving
211 interference when merging models. *arXiv:2306.01708*, 2023.
- 212 [26] Yi-Lin Sung, Linjie Li, Kevin Lin, Zhe Gan, Mohit Bansal, and Lijuan Wang. An empirical
213 study of multimodal model merging, 2023.
- 214 [27] Michael Matena and Colin Raffel. Merging models with fisher-weighted averaging.
215 *arXiv:2111.09832*, 2021.
- 216 [28] C Daniel Freeman and Joan Bruna. Topology and geometry of half-rectified network optimiza-
217 tion. *arXiv:1611.01540*, 2016.
- 218 [29] Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson.
219 Loss surfaces, mode connectivity, and fast ensembling of dnns. *NeurIPS*, 2018.
- 220 [30] Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred Hamprecht. Essentially no
221 barriers in neural network energy landscape. In *ICML*, 2018.
- 222 [31] Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. Linear mode
223 connectivity and the lottery ticket hypothesis. In *ICML*, 2020.
- 224 [32] Norman Tatro, Pin-Yu Chen, Payel Das, Igor Melnyk, Prasanna Sattigeri, and Rongjie Lai.
225 Optimizing mode connectivity via neuron alignment. *NeurIPS*, 2020.
- 226 [33] Sidak Pal Singh and Martin Jaggi. Model fusion via optimal transport. *NeurIPS*, 2020.
- 227 [34] Chang Liu, Chenfei Lou, Runzhong Wang, Alan Yuhan Xi, Li Shen, and Junchi Yan. Deep
228 neural network fusion via graph matching with applications to model ensemble and federated
229 learning. In *ICML*, 2022.
- 230 [35] Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation
231 invariance in linear mode connectivity of neural networks. *arXiv:2110.06296*, 2021.
- 232 [36] Fidel A Guerrero Peña, Heitor Rapela Medeiros, Thomas Dubail, Masih Aminbeidokhti, Eric
233 Granger, and Marco Pedersoli. Re-basin via implicit sinkhorn differentiation. *arXiv:2212.12042*,
234 2022.
- 235 [37] Xiaoxi He, Zimu Zhou, and Lothar Thiele. Multi-task zipping via layer-wise neuron sharing.
236 *NeurIPS*, 2018.
- 237 [38] Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft. Convergent Learning:
238 Do different neural networks learn the same representations? *arXiv:1511.07543*, 2015.

- 239 [39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal,
240 Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual
241 models from natural language supervision. In *ICML*, 2021.
- 242 [40] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for
243 fine-grained image categorization: Stanford dogs. In *CVPR Workshop on Fine-Grained Visual*
244 *Categorization (FGVC)*, 2011.
- 245 [41] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In
246 *CVPR*, 2012.
- 247 [42] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD
248 Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.
- 249 [43] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro
250 Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen
251 scientists: The fine print in fine-grained dataset collection. In *CVPR*, 2015.
- 252 [44] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous
253 convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.
- 254 [45] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual
255 object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June
256 2010.
- 257 [46] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale
258 hierarchical image database. In *CVPR*, 2009.
- 259 [47] Thomas Uriot and Dario Izzo. Safe crossover of neural networks through neuron alignment. In
260 *GECCO*, 2020.
- 261 [48] Aditya Kumar Akash, Sixu Li, and Nicolás García Trillos. Wasserstein barycenter-based model
262 fusion and linear mode connectivity of neural networks. *arXiv:2210.06671*, 2022.
- 263 [49] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep
264 convolutional neural networks. *Communications of the ACM*, 2017.
- 265 [50] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba.
266 Scene parsing through ade20k dataset. In *CVPR*, 2017.
- 267 [51] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing
268 Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset
269 for audio events. In *ICASSP*, 2017.
- 270 [52] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017.
- 271 [53] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object
272 detection. In *ICCV*, 2019.
- 273 [54] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance
274 segmentation. In *ICCV*, 2019.
- 275 [55] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer
276 backbones for object detection. In *ECCV*, 2022.
- 277 [56] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative
278 adversarial networks. In *CVPR*, 2018.
- 279 [57] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer.
280 High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.
- 281 [58] Zhizhong Li and Derek Hoiem. Learning without forgetting. *TPAMI*, 2017.
- 282 [59] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis,
283 Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in
284 classification tasks. *TPAMI*, 2021.

- 285 [60] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen,
286 Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generaliza-
287 tion. *TKDE*, 2022.
- 288 [61] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image
289 recognition. 2016.
- 290 [62] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images.
291 2009.
- 292 [63] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable
293 neural networks. *arXiv:1803.03635*, 2018.
- 294 [64] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai,
295 Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al.
296 An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv:2010.11929*,
297 2020.
- 298 [65] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. Singan: Learning a generative model
299 from a single natural image. In *ICCV*, 2019.
- 300 [66] J MacQueen. Classification and analysis of multivariate observations. In *5th Berkeley Symp.
301 Math. Statist. Probability*, 1967.
- 302 [67] Louis L McQuitty. Elementary linkage analysis for isolating orthogonal and oblique types and
303 typal relevancies. *Educational and psychological measurement*, 1957.
- 304 [68] Wei-Hong Li and Hakan Bilen. Knowledge distillation for multi-task learning. In *ECCV*, 2020.
- 305 [69] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*,
306 2020.
- 307 [70] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transfor-
308 mers. In *CVPR*, 2022.
- 309 [71] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin
310 Gilmer, Andreas Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al.
311 Scaling vision transformers to 22 billion parameters. *arXiv:2302.05442*, 2023.
- 312 [72] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins,
313 Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska,
314 et al. Overcoming catastrophic forgetting in neural networks. *PNAS*, 2017.
- 315 [73] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via
316 invariant feature representation. In *ICML*, 2013.
- 317 [74] Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classifica-
318 tion tasks to a new unlabeled sample. In *NeurIPS*, 2011.
- 319 [75] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images.
320 2009.
- 321 [76] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan,
322 Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*,
323 2014.
- 324 [77] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time
325 object detection with region proposal networks. *NeurIPS*, 2015.
- 326 [78] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics,
327 and function using networkx. In *Proceedings of the 7th Python in Science Conference*, 2008.
- 328 [79] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and
329 Judy Hoffman. Token merging: Your vit but faster. *ICLR*, 2023.

- 330 [80] Hongjoon Ahn, Jihwan Kwak, Subin Lim, Hyeonsu Bang, Hyojun Kim, and Taesup Moon.
331 Ss-il: Separated softmax for incremental learning. In *ICCV*, 2021.
- 332 [81] Berfin Simsek, François Fleuret, Arthur Jacot, Francesco Spadaro, Clement Hongler, Wulfram
333 Gerstner, and Johanni Brea. Geometry of the loss landscape in overparameterized neural
334 networks: Symmetries and invariances. In *Proceedings of the 38th International Conference on
335 Machine Learning*. PMLR, 2021.
- 336 [82] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and
337 generalization in neural networks (invited paper). In *Proceedings of the 53rd Annual ACM
338 SIGACT Symposium on Theory of Computing*, 2021.

339 **A Layer Merging Details**

340 Given M_i and U_i , here we describe rules more concretely for each layer type needed to define most
341 convnets.

342 **Conv.** Apply M_i and U_i to each kernel location (i.e., move the $k \times k$ kernel dimensions to the batch
343 dimension).

344 **BatchNorm.** Apply M_i to all parameters (weight, bias, mean, variance), squaring it for the variance
345 term. Continue propagation. As [3] points out, we cannot compute the correct variance without
346 knowing the covariance between the two models (which we don't have access to). Thus, we reset
347 batch norms after merging to evaluate the variance correctly.

348 **LayerNorm.** Apply M_i to all parameters (weight, bias). Since LayerNorm computes mean and
349 standard deviation on the fly, we don't need to do anything special.

350 **Avg / Max Pool.** Skip.

351 **Skip Connection.** Apply the same M_i to each layer whose outputs are combined via the residual so
352 that their output spaces are all aligned.