A Checkpoints used for Ensembling

For the experiments in section 3.1, we use fine-tuned BERT-base checkpoints from the Hugging Face model hub.[2] Specifically, we used the following checkpoints for each datatset:

RTE

- `textattack/bert-base-uncased-RTE`
- `yoshitomo-matsubara/bert-base-uncased-rte`
- `Ruizhou/bert-base-uncased-finetuned-rte`
- `howey/bert-base-uncased-rte`
- `anirudh21/bert-base-uncased-finetuned-rte`

MRPC

- `textattack/bert-base-uncased-MRPC`
- `yoshitomo-matsubara/bert-base-uncased-mrpc`
- `Maelstrom77/bert-base-uncased-MRPC`
- `Ruizhou/bert-base-uncased-finetuned-mrpc`
- `TehranNLP-org/bert-base-uncased-mrpc-2e-5-42`

SST2

- `aviator-neural/bert-base-uncased-sst2`
- `howey/bert-base-uncased-sst2`
- `yoshitomo-matsubara/bert-base-uncased-sst2`
- `ikevin98/bert-base-uncased-finetuned-sst2`
- `TehranNLP-org/bert-base-uncased-cls-sst2`

B Individual dataset results for robust fine-tuning

Figure 7 shows the individual results from when applying WiSE-FT to five out-of-domain datasets using either isotropic or Fisher merging.

C GLUE Fine-tuning Details

For the high resource tasks QNLI, QQP, SST-2, and MNLI, we used checkpoints downloaded from Hugging Face. We also used a checkpoint from Hugging Face that was fine-tuned on the extractive question answering task SQuAD 2.0 [50] as an alternative intermediate task checkpoint. For the low resource tasks CoLA, MRPC, RTE, and STS-B, we fine-tuned for 10 epochs using a batch size of 16 and the Adam optimizer [25] with a learning rate of 1e-5. We ran 5 independent fine-tuning runs for the low-resource tasks, discarding runs with poor performance.

D Domain-Adaptive Pre-training Details

We performed additional domain-adaptive pre-training on RoBERTa-base for 32,768 steps with a batch size of 32 using the Adam optimizer with a learning rate of 1e-5. We used the BioMed and CS splits of the public S2ORC dataset of abstracts and full-length papers [32]. We note that Gururangan et al. [17] used an internal version of S2ORC that includes additional papers that could not be released due to copyright issues. Our fine-tuning and target task Fisher computation procedures were the same as in our GLUE experiments with the exception of using a batch size of 8 when fine-tuning.

[2] `https://huggingface.co/models`

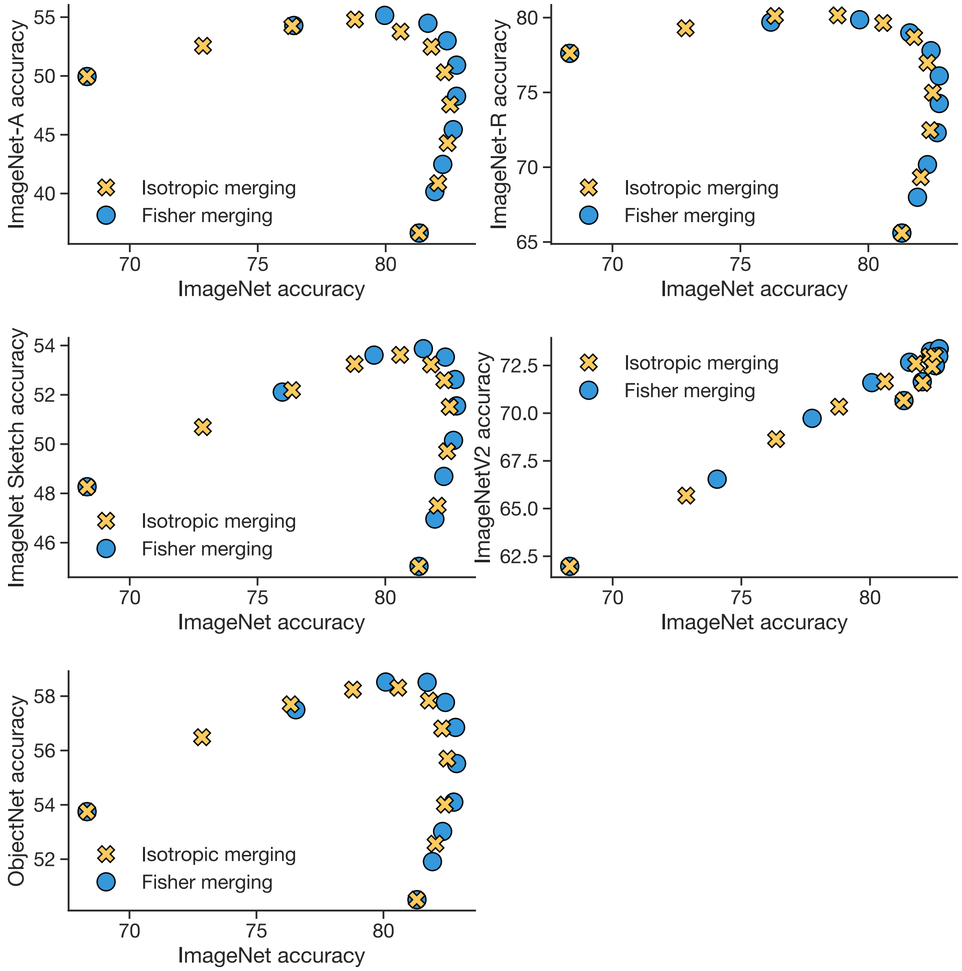

Figure 7: Individual OOD dataset results from applying WiSE-FT [61] to ImageNet pre-trained ViT-B/16 using either isotropic or Fisher merging.

Fine-tuning for 10 epochs, we saved a checkpoint at the end of each epoch. We computed the Fisher for the DAPT checkpoints on 131,072 examples, using one sample from the logits per example. We merged each checkpoint saved during fine-tuning with the DAPT checkpoint from the task's domain. We performed a grid search of 75 merging coefficients and used the F_1 score on the first 2048 test examples as the selection criterion. We report the scores of the best unmerged and the best merged checkpoint from each fine-tuning run.

E Full results for intermediate-task training

In tables A1 to A3, we report results for intermediate-task training when considering all possible datasets in GLUE as target tasks.

F Using fewer examples to estimate the Fisher

In table A4, we show the effect of limiting the number of examples used to compute the Fisher when performing intermediate-task Fisher merging on BERT-base with MNLI as the donor task and RTE as the target task.

Table A1: Intermediate task Fisher merging results on GLUE with BERT-base. Columns correspond to target tasks while rows correspond to intermediate tasks. Italicized values on the diagonal are scores of the unmerged target checkpoints. Subscripts denote standard deviation across runs.

Task	CoLA	SST-2	MRPC	STS-B	QQP	MNLI	QNLI	RTE
CoLA	$\mathit{55.4_{1.8}}$	$92.4_{0.0}$	$84.8_{0.4}$	$86.1_{0.9}$	$89.3_{0.0}$	$83.9_{0.0}$	$90.9_{0.0}$	$64.7_{1.0}$
SST-2	$55.8_{1.6}$	$\mathit{92.4_{0.0}}$	$84.7_{0.5}$	$86.1_{0.9}$	89.0	83.9	90.9	$65.0_{1.4}$
MRPC	$55.7_{1.7}$	$92.4_{0.0}$	$\mathit{84.5_{0.3}}$	$86.1_{0.8}$	$88.9_{0.1}$	$83.8_{0.0}$	$90.9_{0.0}$	$65.0_{0.9}$
STS-B	$55.7_{1.5}$	$92.4_{0.0}$	$84.8_{0.4}$	$\mathit{86.1_{0.8}}$	$88.9_{0.1}$	$83.8_{0.0}$	$90.9_{0.0}$	$65.4_{2.3}$
QQP	$55.5_{1.7}$	92.4	$84.6_{0.3}$	$86.1_{0.9}$	$\mathit{88.8_{0.0}}$	83.8	90.9	$65.8_{2.3}$
MNLI	$55.7_{1.9}$	92.4	$85.1_{0.6}$	$86.1_{0.9}$	88.9	$\mathit{83.7_{0.0}}$	90.9	$73.2_{5.1}$
QNLI	$55.5_{1.7}$	92.4	$85.0_{0.8}$	$86.1_{0.9}$	89.4	83.9	$\mathit{90.9_{0.0}}$	$66.5_{1.6}$
RTE	$55.6_{1.7}$	$92.4_{0.0}$	$84.8_{0.4}$	$86.1_{0.9}$	$88.8_{0.0}$	$83.9_{0.1}$	$90.9_{0.0}$	$\mathit{63.7_{1.7}}$
SQuAD	$56.1_{1.4}$	92.4	$84.9_{0.4}$	$86.1_{0.9}$	89.1	83.9	91.0	$66.6_{1.0}$

Table A2: Intermediate task isotropic-merging results on GLUE with BERT-base. Columns correspond to target tasks while rows correspond to intermediate tasks. Italicized values on the diagonal are scores of the unmerged target checkpoints. Subscripts denote standard deviation across runs.

Task	CoLA	SST-2	MRPC	STS-B	QQP	MNLI	QNLI	RTE
CoLA	$\mathit{55.4_{1.8}}$	$92.5_{0.0}$	$84.9_{0.8}$	$86.1_{0.8}$	$88.9_{0.0}$	$83.9_{0.0}$	$90.9_{0.0}$	$64.8_{0.8}$
SST-2	$55.5_{1.7}$	$\mathit{92.4_{0.0}}$	$84.9_{0.7}$	$86.1_{0.9}$	88.8	83.9	90.9	$64.8_{1.1}$
MRPC	$55.5_{1.8}$	$92.4_{0.0}$	$\mathit{84.5_{0.3}}$	$86.1_{0.9}$	$88.9_{0.1}$	$83.9_{0.1}$	$90.9_{0.0}$	$65.1_{0.9}$
STS-B	$55.4_{1.8}$	$92.4_{0.0}$	$85.0_{0.4}$	$\mathit{86.1_{0.9}}$	$89.0_{0.1}$	$83.8_{0.1}$	$90.9_{0.0}$	$65.2_{2.1}$
QQP	$55.5_{1.8}$	92.4	$84.7_{0.2}$	$86.1_{0.9}$	$\mathit{88.8_{0.0}}$	83.8	90.9	$65.1_{1.7}$
MNLI	$55.6_{1.7}$	92.4	$85.4_{0.6}$	$86.1_{0.8}$	88.8	$\mathit{83.7_{0.0}}$	90.9	$72.2_{4.0}$
QNLI	$55.5_{1.7}$	92.4	$85.1_{0.7}$	$86.1_{0.9}$	89.1	83.9	$\mathit{90.9_{0.0}}$	$66.8_{1.1}$
RTE	$55.5_{1.8}$	$92.4_{0.0}$	$84.6_{0.3}$	$86.1_{0.8}$	$88.9_{0.1}$	$83.8_{0.1}$	$90.9_{0.0}$	$\mathit{63.7_{1.7}}$

Table A3: Sequential fine-tuning results on GLUE with BERT-base. Columns correspond to target tasks while rows correspond to intermediate tasks. Subscripts denote standard deviation across runs. Italicized values represent fine-tuning directly on the target task (i.e. no intermediate-task training).

Task	CoLA	MRPC	STS-B	RTE
CoLA	$\mathit{55.4_{1.8}}$	$85.0_{0.9}$	$85.9_{0.8}$	$62.1_{2.3}$
SST-2	$56.8_{1.4}$	$85.4_{0.9}$	$85.3_{1.0}$	$63.8_{1.0}$
MRPC	$58.5_{0.4}$	$\mathit{84.5_{0.3}}$	$85.3_{0.8}$	$62.7_{5.2}$
STS-B	$56.3_{0.4}$	$86.7_{0.7}$	$\mathit{86.1_{0.9}}$	$64.5_{2.5}$
QQP	$56.0_{2.0}$	$87.1_{1.2}$	$87.5_{0.4}$	$71.6_{1.9}$
MNLI	$58.6_{1.7}$	$85.9_{0.8}$	$87.6_{0.3}$	$77.4_{1.6}$
QNLI	$56.4_{1.9}$	$87.8_{0.6}$	$87.1_{0.5}$	$71.0_{4.1}$
RTE	$56.7_{0.9}$	$82.2_{2.5}$	$85.8_{0.5}$	$\mathit{63.7_{1.7}}$

Table A4: Effect of the number of examples used to compute the Fisher information. Columns correspond to the number of examples used for RTE. Rows correspond to the number of examples used for MNLI. Scores are the RTE validation set accuracy. The original RTE checkpoints had an average accuracy of 63.7 and isotropic merging (i.e. 0 Fisher examples) had an average accuracy of 72.2.

Examples	256	1024	2490
256	72.7	72.9	73.1
1024	72.9	72.9	73.3
4096	72.9	73.0	73.2
32768	72.8	73.0	73.5
392702	72.9	73.1	73.4