1 Pruning for GPU speedups (N:M sparsity pattern)

Table 1: Comparison between oBERT one-shot 2-out-of-4 (2:4) pruning and Magnitude Pruning. Methods such as Lottery-Ticket, Movement Pruning, Prune OFA, l_0 Regularization, and PLATON require fine-tuning and therefore do not support one-shot pruning. Fine-tuning the oBERT 2:4 pruned model for only 1-epoch fully recovers dense model accuracy with (F1, EM) = (88.58, 81.16).

Task	BERT-Base	Magnitude	oBERT (ours)	GPU speedup
SQuAD F1, EM	88.54, 81.41	49.97, 35.24	**83.17, 74.18**	1.85x

2 Additional models: unstructured pruning

Table 2: **Comparison between oBERT and the upstream SOTA Prune OFA method, when pruning the BERT-Large model at 90% sparsity**. Even the model pruned with oBERT at double the sparsity (95%) outperforms Prune OFA.

Task	BERT-Large	Sparsity	Prune OFA	oBERT (ours)
SQuAD F1, EM	91.22, 84.45	90%	90.20, 83.35	**91.00, 84.50**
SQuAD F1, EM	91.22, 84.45	95%	NA	90.29, 83.58

3 Additional models: compound compression for edge deployment

Table 3: **Compressing BERT-Large and MobileBERT models on the SQuADv1 task, with the goal of recovering >99% of the dense BERT-Large accuracy**. oBERT-Large stands for our 95% block4 pruned and quantized BERT-Large model, and oBERT-MobileBERT stands for a 14-layer, 50% block4 pruned and quantized MobileBERT model. Both models are produced following the compound compression approach described in the paper. Models were evaluated with the DeepSparse inference engine, using a server with two Intel(R) Xeon(R) Platinum 8380 (IceLake) CPUs with 40 cores each, batch-size 128 and sequence length 384.

Model	Precision	F1 Score (R=X% recovery)	File Size	Compression Ratio	Throughput (samples/sec)	Speedup
BERT-Large dense baseline	FP32	90.87 (R=100%)	1.30 GB	1x	15.49	1x
oBERT-Large	INT8	90.21 (R=99.27%)	38.20 MB	34x	230.74	15x
oBERT-MobileBERT	INT8	90.32 (R=99.39%)	9.56 MB	136x	928.58	60x

4 Additional GLUE results

Table 4: **Additional results on GLUE tasks where baselines from other work are available, complementing Table 2 in the submission**. By contrast to competing work, oBERT results are obtained **without any per-task hyper-parameter tuning**. The same setup is used for all results presented in the paper (SQuAD, MNLI, QQP) and results shown here (SST-2, QNLI).

Task	BERT-Base	Sparsity	LT-BERT	Prune OFA	oBERT (ours)
SST-2 Accuracy	93.01	90%	85.00*	90.88	**92.20**
QNLI Accuracy	91.25	90%	80.00*	89.07	**89.97**

5 Additional GLUE results + comparison with *concurrent* work

Table 5: **Comparison with concurrent work PLATON (ICML 2022)**. For fair comparison, we remove Knowledge-Distillation (KD) during fine-tuning because the competing methods do not use it. All oBERT results are obtained **without any per-task hyper-parameter tuning**, except for early stopping to prevent overfitting on tiny GLUE tasks. By contrast, the competing PLATON work reports best results after extensive task-specific hyper-parameter search. The results are reported at 90%, the highest sparsity target in the PLATON work, and *NA* indicates the model does not converge. The best-performing method is marked in green.

Task	BERT BASE	l_0 Regularization	Magnitude	Movement	Soft-Movement	PLATON	oBERT (ours)
MNLI m / mm	84.6 / 83.4	78.0 / 78.7	78.8 / 79.0	79.3 / 79.5	80.7 / 81.1	82.0 / 82.2	**82.2 / 82.5**
QQP Acc / F1	91.5 / 88.5	87.6 / 82.0	78.8 / 77.0	89.1 / 85.4	90.2 / 86.7	90.2 / 86.8	**90.4 / 87.1**
QNLI Acc	91.3	82.8	86.6	79.2	86.6	88.9	**89.3**
MRPC Acc / F1	86.4 / 90.3	73.8 / 79.5	70.3 / 80.3	68.4 / 81.2	79.7 / 85.9	84.3 / 88.8	**85.6 / 89.3**
SST-2 Acc	92.7	82.5	80.7	80.2	87.4	90.5	**92.0**
CoLA Mcc	58.3	NA	NA	NA	NA	44.3	**48.47**
STS-B Pear / Spear	90.2 / 89.7	82.7 / 83.9	83.4 / 83.3	NA	86.5 / 86.3	87.4 / 87.1	**88.0 / 87.6**