A Distributions of Rank

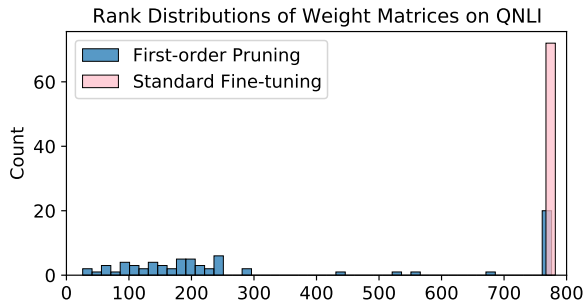

Figure 1: The rank distribution of weight matrices of models trained by standard fine-tuning and first-order pruning (SMvP). For zero-order pruning (both post-training one-shot magnitude pruning and Lotter Ticket Hypothesis), we observe a similar distribution as standard fine-tuning, hence we do not show it in the figure.

As described in the pilot experiment, the average rank of models produced by first-order pruning algorithms is much lower than that of models produced by standard fine-tuning and zero-order pruning algorithms. Concretely, we show the respective rank distribution on QNLI dataset in Figure 1. As can be observed, for first-order pruning, a large portion of ranks are distributed at 100-200. In stark contrast, the densely fine-tuned model is nearly full-rank. This phenomenon highlights the possibility of effective model compression using low-rank factorization on models produced by first-order pruning algorithms.

B Low-rank Factorization on Zero-order Pruning Algorithms

In the paper, we primarily adopt soft-movement pruning (SMvP) as our default choice of first-order pruning algorithm. To give more clear evidence of how the rank-distribution of a model affects the downstream performance after factorization, we also present the results of post-training one-shot magnitude pruning (POMP) and lottery ticket hypothesis (LTH) in Table 2 and Table 3. Compared to Table 1 in Section 3, we can see that the downstream performance of low-rank factorization upon models produced by zero-order pruning algorithms is on par with that of SVD while being much lower than our proposed LPAF, verifying the necessity of a low-rank structure for low-rank factorization to retain satisfactory task performance.

C Detailed Results of Compared Compression Methods

We present the detailed experimental results of all compared BERT compression methods in Table 1. All these methods are able to offer a perceivable reduction in memory and computation without resorting to specialized linear algebra implementation or hardware. We can categorize these methods into four types:

- Task-agnostic distillation methods, which include DistilBERT, PD-BERT, and TinyBERT. They realize different levels of compression rate by varying the number of encoder layers.

- Task-specific distillation methods, which include baselines in the second part of Table 1. They realize different levels of compression rate by varying the number of encoder layers They also differ from how the student model is initialized: PKD initializes the student model with the lowest fewer layers from the teacher model; RAIL-KD leverage DistilBERT as initialization for distillation, which naturally provides a better starting point. CKD instead uses PD-BERT as initialization, which is also proven to be more effective than the strategy of PKD.

- We also include the results of Iterative Structured Pruning (**ISP**) which progressively removes attention heads in multi-head attention layer and neurons in the feed-forward layer with the lowest importance scores. The importance score of a specific structure s is measured by the expected gradient of the loss function $\mathcal{L}$ with respect to the mask variable ϵ_s associated with it:

$$I_s = \mathbb{E}_{x\sim\mathbb{X}}|\frac{\partial\mathcal{L}(x)}{\partial\epsilon_s}| \quad (1)$$

ISP achieves BERT compression by locally slimming specific structures without explicitly reducing number of encoder layers. Different from ISP that reach target compression rate in an iterative manner, we experiment with structured pruning with l_0 regularization that explicitly control the compression rate by modifying the training objective, denoted as **SP-l_0**. SP-l_0 draws samples from the hard-concrete distribution as approximation to binary masks for different structures.

Models	SST-2 (67K)			QQP (364K)			QNLI (105K)			MNLI-m (393K)			SQuAD v1.1 (88K)		
Metric	Accuracy			Accuracy			Accuracy			Accuracy			F1 score		
Percent of #Params ↓	25%	16%	8%	25%	16%	8%	25%	16%	8%	25%	16%	8%	25%	16%	8%
Task-agnostic Distillation															
DistilBERT	88.9	86.4	83.0	89.4	88.0	82.4	83.8	81.6	64.9	76.4	71.6	59.8	78.0	66.5	28.5
PD-BERT	88.2	87.5	82.7	89.8	88.9	82.9	86.1	83.4	65.9	78.6	75.9	66.0	77.0	45.2	22.8
TinyBERT	89.8	88.0	82.6	90.0	88.7	83.2	87.7	84.5	63.1	80.6	77.1	65.6	58.0	38.1	15.4
Task-specific Distillation															
$\text{BERT}_{\text{Truncated}}$	87.7	85.6	79.6	88.0	86.9	82.2	83.0	78.7	60.3	75.7	73.1	60.6	54.8	31.4	17.1
Vanilla KD	88.5	86.5	82.0	88.8	87.2	82.3	83.6	78.6	59.8	76.1	73.3	61.3	62.5	32.3	17.0
PKD	88.1	87.2	82.0	88.5	87.5	82.3	82.7	78.0	59.8	75.8	73.0	61.3	60.5	31.5	17.0
BERT-of-Theseus	88.5	86.1	79.6	89.0	86.0	82.2	85.0	80.3	60.3	76.3	73.4	60.6	72.7	63.2	26.2
RAIL-KD	88.8	86.8	83.9	90.2	88.6	83.3	85.7	81.2	65.7	78.8	73.5	62.1	78.7	69.5	32.1
CKD	89.8	88.7	84.1	90.1	88.9	82.9	87.0	84.9	67.6	79.1	76.8	66.4	78.9	69.1	33.1
MetaDistil	88.9	87.0	82.7	88.9	86.9	82.1	86.8	84.9	67.5	79.2	76.7	66.4	78.8	69.0	32.4
Model architecture-dependent structured pruning															
SP-l_0	90.4	89.4	87.9	90.1	89.3	85.5	88.7	87.2	84.9	81.9	80.8	76.2	84.9	81.9	69.8
ISP	89.7	89.2	87.0	90.2	89.9	89.0	88.6	88.1	84.9	80.8	80.5	78.5	81.7	78.7	73.7
Factorization-based															
SVD	88.9	88.1	84.5	90.0	87.9	83.1	86.1	83.8	67.6	79.6	76.6	71.6	85.5	81.1	51.3
LPAF-CAP (ours)	89.8	89.2	87.5	**90.5**	**90.4**	**89.3**	89.0	88.3	85.2	81.9	**81.5**	**79.5**	86.5	85.6	81.8
LPAF-SMvP (ours)	**90.7**	**89.7**	**88.5**	90.4	90.1	89.0	**89.2**	**88.6**	**85.7**	**82.2**	81.4	79.2	**87.2**	**85.7**	**82.0**

Table 1: Experimental resutls of all compared BERT compression methods. The best results are **bolded** and the second best are underlined. The numbers in the parenthesis are the size of training data for each task.

Model	SST-2	QNLI	QQP
POMP-100	87.50	84.72	89.11
POMP-80	86.12	83.07	88.32
POMP-60	85.89	80.29	87.16
POMP-40	84.29	70.49	83.47

Table 2: Experimental results of low-rank factoriztion upon models produced by post-training one-shot magnitude pruning (POMP).

Model	SST-2	QNLI	QQP
LTH-100	87.50	84.50	89.10
LTH-80	86.70	83.86	88.46
LTH-60	85.00	79.81	87.54
LTH-40	85.67	68.00	85.77

Table 3: Experimental results of low-rank factoriztion upon models produced by the lottery ticket hypothesis (LTH).

Note that we omit ISP and SP-l_0 in the main experiment because they are both model architecture-dependent and require troublesome re-organization of different model structures to realize perceptible memory and computation reduction. In contrast, LPAF performs *inerratic* matrix decomposition, making efficient inference straightforward.

- Factorization-based methods including SVD and our LPAF that shrink the original model without changing the number of layers.

Although many task-specific compression methods like BERT-of-Theseus have demonstrated improvement over task-agnostic methods like TinyBERT under conventional compression setting (e.g., compressing a 12-layer teacher model into a 6-layer student model), they actually perform worse under relatively high compression ratio, i.e., $\geq$ 4.0x. The reason might be that under a high compression regime, pure task-specific compression without utilizing the pre-compression stage has limited generalization ability due to: 1) insufficient network capacity. 2) ineffective knowledge transfer from the teacher because of the large capacity gap in between.