Extreme Compression for Pre-trained Transformers Made Simple and Efficient

Anonymous Author(s)
Affiliation
Address
email

Abstract

Extreme compression, particularly ultra-low bit precision (binary/ternary) quantization, has been proposed to fit large NLP models on resource-constraint devices. However, to preserve the accuracy for such aggressive compression schemes, cutting-edge methods usually introduce complicated compression pipelines, e.g., multi-stage expensive knowledge distillation with extensive hyperparameter tuning. Also, they oftentimes focus less on smaller transformer models that have already been heavily compressed via knowledge distillation and lack a systematic study to show the effectiveness of their methods. In this paper, we perform a very comprehensive systematic study to measure the impact of many key hyperparameters and training strategies from previous works. As a result, we find out that previous baselines for ultra-low bit precision quantization are significantly under-trained. Based on our study, we propose a simple yet effective compression pipeline for extreme compression, named XtrmC. XtrmC demonstrates that (1) we can skip the pre-training knowledge distillation to obtain a 5-layer BERT while achieving better performance than previous state-of-the-art methods, e.g., the 6-layer TinyBERT; (2) extreme quantization plus layer reduction is able to reduce the model size by 50x, resulting in new state-of-the-art results on GLUE tasks.

1 Introduction

Over the past few years, we have witnessed the model size has grown at an unprecedented speed, from a few hundred million parameters (e.g., BERT [8], RoBERTa [26], DeBERTA [12], T5 [35],GPT-2 [34]) to a few hundreds of billions of parameters (e.g., 175B GPT-3 [6], 530B MT-NLT [43]), showing outstanding results on a wide range of language processing tasks. Despite the remarkable performance in accuracy, there have been huge challenges to deploying these models, especially on resource-constrained edge or embedded devices. Many research efforts have been made to compress these huge transformer models including knowledge distillation [18, 50, 49, 58], pruning [52, 40, 7], and low-rank decomposition [28]. Orthogonally, quantization focuses on replacing the floating-point weights of a pre-trained Transformer network with low-precision representation. This makes quantization particularly appealing when compressing models that have already been optimized in terms of network architecture.

Popular quantization methods include post-training quantization [42, 31, 27], quantization-aware training (QAT) [21, 4, 29, 20], and their variations [41, 22, 9]. The former directly quantizes trained model weights from floating-point values to low precision values using a scalar quantizer, which is simple but can induce a significant drop in accuracy. To address this issue, quantization-aware training directly quantizes a model during training by quantizing all the weights during the forward and using a straight-through estimator (STE) [5] to compute the gradients for the quantizers.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

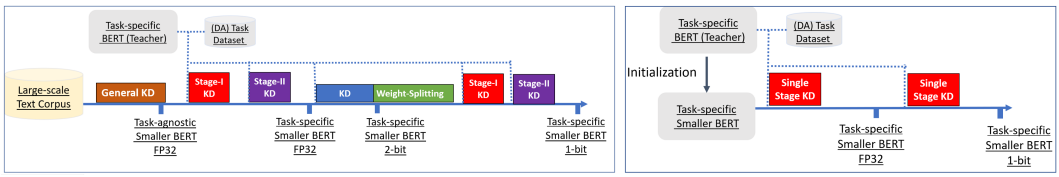

Figure 1: The left figure summarizes how to do 1-bit quantization for a layer-reduced model based on [18, 3]. It involves expensive pretraining on an fp-32 small model, task-specific training on 32-bit and 2-bit models, weight-splitting, and the final 1-bit model training. Along the way, it applies multi-stage knowledge distillation with data augumentation, which needs considerable hyperparameter tuning efforts. The right figure is our proposed method, XtrmC (see details in § 5), a simple while effective pipeline (see Figure 2 for highlighted results). Better read with a computer screen.

Recently, several QAT works further push the limit of BERT quantization to the extreme via ternarized (2-bit) ([56]) and binarized (1-bit) weights ([3]) together with 4-/8-bit quantized activation. These compression methods have been referred to as **extreme quantization** since the limit of weight quantization, in theory, can bring over an order of magnitude compression rates (e.g., 16-32 times). One particular challenge identified in [3] was that it was highly difficult to perform binarization as there exits a sharp performance drop from ternarized to binarized networks. To address this issue, prior work proposed multiple optimizations where one first trains a ternarized DynaBert [15] and then binarizes with weight splitting. In both phases, multi-stage distillation with multiple learning rates tuning and data augmentation [18] are used. Prior works claim these optimizations are essential for closing the accuracy gap from binary quantization.

While the above methodology is promising, several unanswered questions are related to recent extreme quantization. First, as multiple ad-hoc optimizations are applied at different stages, the compression pipeline becomes very complex and expensive, limiting the applicability of extreme quantization in practice. Moreover, a systematical evaluation and comparison of these optimizations are missing, and the underlying question remains open for extreme quantization:

what are the necessities of ad-hoc optimizations to recover the accuracy los?

Second, prior extreme quantization primarily focused on reducing the precision of the network. Meanwhile, several advancements have also been made in the research direction of knowledge distillation, where large teacher models are used to guide the learning of a small student model. Examples include DistilBERT [39], MiniLM [50, 49], MobileBERT [45], which demonstrate 2-4× model size reduction by reducing the depth or width without much accuracy loss through pre-training distillation with an optional fine-tuning distillation. Notably, TinyBERT [18] proposes to perform deep distillation in both the pre-training and fine-tuning stages and shows that this strategy achieves state-of-the-art results on GLUE tasks. However, most of these were done without quantization, and there are few studies about the interplay of extreme quantization with these heavily distilled models, which poses questions on

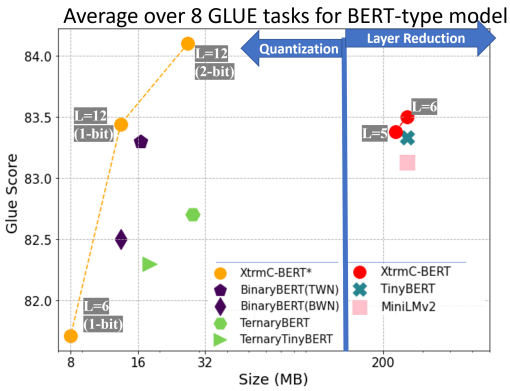

Figure 2: The comparison between XtrmC with other SOTA results.

to what extend can smaller distilled models benefit from extreme quantization?

Contribution. To investigate the above questions, we make the following contributions: (1) We present a systematic study of extreme quantization methods by fine-tuning ≥1000 pre-trained Transformer models, which includes a careful evaluation of the effects of hyperparameters and several methods introduced in extreme quantization. (2) We find that previous extreme quantization studies overlooked certain design choices, which lead to under-trained binarized networks and unnecessarily complex optimizations. Instead, we derive a celebrating recipe for extreme quantization, which is

not only simpler but also allows us to achieve an even larger compression ratio and higher accuracy than existing methods (see Figure 1, right). (3) We find that extreme quantization can be effectively combined with lightweight layer reduction, which allows us to achieve greater compression rates for pre-trained Transformers with better accuracy than prior methods while enjoying the additional benefits of flexibly adjusting the size of the student model for each use-case individually, without the expensive pre-training distillation.

Our evaluation results, as illustrated in Figure 2, show that our simple yet effective method can: (1) compress $\text{BERT}_{\text{base}}$ to a 5-layer $\text{BERT}_{\text{base}}$ while achieving better performance than previous state-of-the-art distillation methods, e.g., the 6-layer TinyBERT [18], without incurring the computationally expensive pre-training distillation; and (2) reduce the model size by $50\times$ via performing robust extreme quantization on lightweight layer reduced models while obtaining better accuracy than prior extreme quantization methods, e.g., the 12-layer 1-bit $\text{BERT}_{\text{base}}$, resulting in new state-of-the-art results on GLUE tasks.

2 Related Work

Quantization becomes practically appealing as it not only reduces memory bandwidth consumption but also brings the additional benefit of further accelerating inference on supporting hardware [41, 22, 9]. Most quantization works focus on INT8 quantization [19] or mixed INT8/INT4 quantization [41]. Our work differs from those in that we investigate extreme quantization where the weight values are represented with only 1-bit or 2-bit at most. There are prior works that show the feasibility of using only ternary or even binary weights [56, 3]. Unlike those work, which uses multiple optimizations with complex pipelines, our investigation leads us to introduce a simple yet more efficient method for extreme compression with more excellent compression rates.

On a separate line of research, reducing the number of parameters of deep neural networks models with full precision (no quantization) has been an active research area by applying the powerful knowleadge distillation (KD) [14, 50], where a stronger teacher model guides the learning of another small student model to minimize the discrepancy between the teacher and student outputs. Please see Appendix A for a comprehensive literature review on KD. Instead of proposing a more advanced distillation method, we perform a well-rounded comparative study on the effectiveness of the recently proposed multiple-stage distillation [18].

3 Extreme Compression Procedure Analysis

This section presents several studies that have guided the proposed method introduced in Section 5. All these evaluations are performed with the General Language Understanding Evaluation (GLUE) benchmark [47], which is a collection of datasets for evaluating natural language understanding systems. For the subsequent studies, we report results on the development sets after compressing a pre-trained model (e.g., $\text{BERT}_{\text{base}}$ and TinyBERT) using the corresponding single-task training data. Previous works [56, 3] on extreme quantization of transformer models state three hypotheses for what can be related to the difficulty of performing extreme quantization:

- Directly training a binarized BERT is complicated due to its irregular loss landscape, so it is better to first to train a ternarized model to initialize the binarized network.
- Specialized distillation that transfers knowledge at different layers (e.g., intermediate layer) and multiple stages is required to improve accuracy.
- Small training data sizes make extreme compression difficult.

3.1 Is staged ternary-binary training necessary to mitigate the sharp performance drop?

Previous works demonstrate that binary networks have more irregular loss surface than ternary models using curvature analysis. However, given that rough loss surface is a prevalent issue when training neural networks, we question whether all of them have a causal relationship with the difficulty in extreme compression.

We are first interested in the observation from previous work [3] that shows directly training a binarized network leads to a significant accuracy drop (e.g., up to 3.8%) on GLUE in comparison to

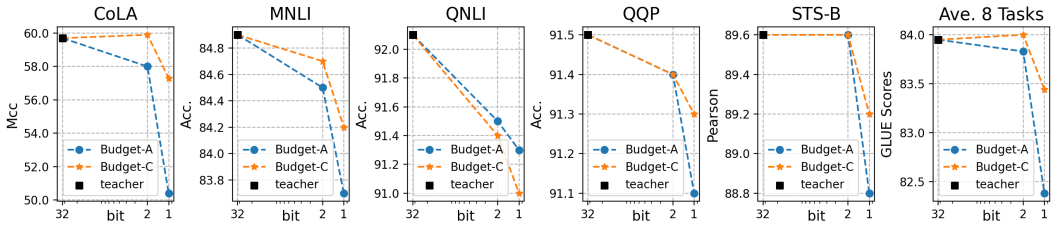

Figure 3: Performance of quantized $\text{BERT}_{\text{base}}$ with different weight bits and 8-bit activation on the GLUE Benchmarks. The results for orange and blue curves respectively represent the costs: (limited) Budget-A and (sufficient) Budget-C. The fp32-teacher scores are shown by black square marker.

the accuracy loss from training a ternary network (e.g., up to 0.6%). They suggest that this is because binary networks have a higher training loss and overall more complex loss surface. To overcome this sharp accuracy drop, they propose a staged ternary-binary training strategy in which one needs first to train a ternarized network and use a technique called weight splitting to use the trained ternary weights to initialize the binary model.

It is not surprising that binarization leads to further accuracy drop as prior studies from computer vision also observe similar phenomenon [16, 46]. However, prior works also identified that the difficulty of training binarized networks mostly resides in training with insufficient number of iterations or having too large learning rates, which can lead to frequent sign changes of the weights that make the learning of binarized networks unstable [46]. Therefore, training a binarized network the same way as training a full precision network consistently achieves poor accuracy. Would increasing the training iterations and let the binarized network train longer under smaller learning rates help mitigate the performance drop from binarization?

To investigate this, we perform the following experiment: We remove the TernaryBERT training stage and weight splitting and directly train a binarized network. We use the same quantization schemes as in [3], e.g., a binary quantizer to quantize the weights: $w_i = \alpha \cdot sign(w_i), \alpha = \frac{1}{n}||w||_1$, and a uniform symmetric INT8 quantizer to quantize the activations. We apply the One-Stage quantization-aware KD that will be explained in § 3.2 to train the model. Nevertheless, without introducing the details of KD here, it does not affect our purpose of understanding the phenomena of a sharp performance drop since we only change the training iteration and learning rates while fixing other setups.

We consider three budgets listed in Table 1, which cover the practical scenarios of short, standard, and long training time, where bucket-A and Budget-B take into account the training details that appeared in BinaryBERT [3], TernaryBERT [56]. Bucket-c has a considerably larger training budget but smaller than TinyBert [3]. Meanwhile, we also perform a grid search of peak learning rates {2e-5, 1e-4, 5e-4}. For more training details on iterations and batch size per iteration, please see Table C.1.

Table 2: **1-bit** quantization for $\text{BERT}_{\text{base}}$ with various Budget-A, Budget-B and Budget-C.

#	Cost	CoLA Mcc	MNLI-m/-mm Acc/Acc	MRPC F1/Acc	QNLI Acc	QQP F1/Acc	RTE Acc	SST-2 Acc	STS-B Pear/Spea	Avg.	Acc Drop
0	Teacher	59.7	84.9/85.6	90.6/86.3	92.1	88.6/91.5	72.2	93.2	90.1/89.6	83.95	-
1	Budget-A	50.4	83.7/84.6	90.0/85.8	91.3	88.0/91.1	72.9	92.8	88.8/88.4	82.38	1.57
2	Budget-B	55.6	84.1/84.4	90.4/86.0	90.8	88.3/91.3	72.6	93.1	88.9/88.5	82.98	0.97
3	Budget-C	57.3	84.2/84.4	90.7/86.5	91.0	88.3/91.3	74.0	93.1	89.2/88.8	83.44	0.51

Table 1: Different training budgets for the GLUE tasks.

Dataset	Data Aug.	Training epochs: Budget-A	Budget-B	Budget-C
QQP/MNLI	✗	3	9	18 or 36
QNLI	✓	1	3	6 or 9
SST-2/STS-B/RTE	✓	1	3	12
CoLA/MRPC	✓	1	3	12 or 18

We present our main results in Table 2 and Figure 3 (see Table C.2 a full detailed results including three learning rates). We observe that although training binarized BERT with more iterations does not fully close the accuracy gap to the uncompressed full-precision model, the sharp accuracy drop from ternarization to binarization (the blue curve v.s. the orange curve in Figure 3) has largely been mitigated, leaving a much smaller accuracy gap to close. For example, when increasing the training time from Budget-A to Budget-C, the performance

of CoLA boosts from 50.4 to 57.3, and the average score improves from 82.38 to 83.44 (Table 2). These results indicate that previous studies [3] on binarized BERT were significantly under-trained. It also suggests that the observed sharp accuracy drop from binarization is primarily an optimization problem. We note that in the context of fine-tuning BERT models, Mosbach et al. [30] also observe that fine-tuning BERT can suffer from vanishing gradients, which causes the training converges to a "bad" valley with sub-optimal training loss. As a result, the authors also suggest increasing the number of iterations to train the BERT model to have stable and robust results.

> **Finding 1.** A longer training iterations with learning rate decay is highly preferred for closing the accuracy gap of extreme quantization.

Finding 1 seems natural. However, we argue that special considerations need to be taken for effectively doing extreme quantization before resorting to more complex solutions.

3.2 The role of multi-stage knowledge distillation

The above analysis shows that the previous binarized BERT models were severely undertrained. In this section, we further investigate the necessity of multi-stage knowledge distillation, which is quite complex because each stage has its own set of hyperparameters. Still, prior works claim to be crucial for improving the accuracy of extreme quantization. Investigating this problem is interesting because it can potentially admit a more straightforward solution with a cheaper cost.

Prior work proposes an interesting knowledge distillation (KD), and here we call it *Two-Stage* KD (2S-KD) [18], which has been applied for extreme quantization [3]. The 2S-KD minimizes the losses of hidden states $\mathcal{L}_{\text{hidden}}$ and attention maps $\mathcal{L}_{\text{att}}$ and losses of prediction logits $\mathcal{L}_{\text{logit}}$ in two separate steps, where different learning rates are used in these two stages, e.g., a $\times 2.5$ larger learning rate is used for Stage-I than that for Stage-II. However, how to decide the learning rates and training epochs for these two stages is not well explained and will add additional hyperparameter tuning costs. We note that this strategy is very different from the deep knowledge distillation strategy used in [44, 49, 45], where knowledge distillation is performed with a single stage that minimizes the sum of the losses from prediction logits, hidden states, and attention maps. Despite the promising results in [3], the mechanism for why multi-stage KD improves accuracy is not well understood.

Generally, the knowledge distillation for transformer models can be formulated as minimizing the following objective:

$$\min_{\theta} \mathbb{E}_{x\sim D}[\gamma\mathcal{L}_{\text{logit}}(x;\theta) + \beta(\mathcal{L}_{\text{att}}(x;\theta) + \mathcal{L}_{\text{hidden}}(x;\theta)] \tag{1}$$

where $\mathcal{L}_{\text{logit}}$ denote the loss (e.g., KL divergence or mean square error) between student's and teacher's prediction logits, and $\mathcal{L}_{hidden}$ and $\mathcal{L}_{\text{att}}$) measures the loss of hidden states and attention maps. $\gamma \in \{0, 1\}, \beta \in \{0, 1\}$ are hyperparameters. See the detailed mathematical definition in § B

To investigate the effectiveness of multi-stage, we compare three configurations (shown in Figure 4):

(1) **1S-KD (One-stage KD):** $(\gamma, \beta) = (1, 1)$ for $t \leq \mathcal{T}$;

(2) **2S-KD (Two-stage KD):** $(\gamma, \beta) = (0, 1)$ if $t < \mathcal{T}/2$; $(\gamma, \beta) = (1, 0)$ if $\mathcal{T}/2 \leq t \leq \mathcal{T}$;

(3) **3S-KD (Three-stage KD):** $(\gamma, \beta) = (0, 1)$ if $t < \mathcal{T}/3$; $(\gamma, \beta) = (1, 1)$ if $\mathcal{T}/3 \leq t \leq 2\mathcal{T}/3$; $(\gamma, \beta) = (1, 0)$ if $2\mathcal{T}/3 \leq t \leq \mathcal{T}$.

where $\mathcal{T}$ presents the total training budget and t represents the training iterations.

Notably, we created the 3S-KD, where we add a transition stage: $\mathcal{L}_{\text{hidden}} + \mathcal{L}_{\text{att}} + \mathcal{L}_{\text{logit}}$ in between the first and second stage of the 2S-KD. The idea is to make the transition of the training objective smoother in comparison to the 2S-KD. In this case, the learning rate schedules will be correspondingly repeated in three times, and the peak learning rate of Stage-I is $\times 2.5$ higher than that of Stage-II and Stage-III. We apply all three KD methods independently (under a fixed random seed) for binary quantization across different learning rates {2e-5, 1e-4, 5e-4} under the same training budget. We do not tune any other hyperparameters.

Table 3 shows the results under Budget-A (see Table C.2 for Budget-B/C). We note that when the learning rate is fixed to 2e-5 (i.e., Row 1, 5, and 9), the previously proposed 2S-KD [18] (81.46)

does show performance benefits over 1S-KD (80.33). Our newly created 3S-KD (81.57) obtains even better accuracy under the same learning rate. However, multi-stage makes the extreme compression procedure complex and inefficient because it needs to set the peak learning rates differently for different stages. Surprisingly, the simple 1S-KD easily outperforms both 2S-KD and 3S-KD (Row 6 and 8) when we slightly increase the search space of the learning rates (e.g., 1e-4 and 5e-4). This means we can largely omit the burden of tuning learning rates and training epochs for each stage in multi-stage by using single-stage but with the same training iteration budget.

> **Finding 2.** Single-stage knowledge distillation with more training budgets and search of learning rates is sufficient to match or even exceed accuracy from multi-stage ones.

Table 3: **1-bit** quantization for $\text{BERT}_{\text{base}}$ with three different KD under the training Budget-A.

#	Stages	learning rate	CoLA Mcc	MNLI-m/-mm Acc/Acc	MRPC F1/Acc	QNLI Acc	QQP F1/Acc	RTE Acc	SST-2 Acc	STS-B Pear/Spea	Avg. all
1	One-Stage	2e-5	44.6	83.1/83.7	88.8/83.1	91.1	87.4/90.7	66.1	92.8	87.8/87.5	80.33
2		1e-4	50.4	83.7/84.6	90.1/85.3	91.3	88.0/91.1	71.5	92.7	88.8/88.4	82.16
3		5e-4	42.3	83.3/84.1	90.0/85.8	89.5	87.8/90.9	72.9	92.5	88.1/87.8	81.04
4		Best (above)	50.4	83.7/84.6	90.0/85.8	91.3	88.0/91.1	72.9	92.8	88.8/88.4	**82.38**
5	Two-Stage	2e-5	48.2	83.3/83.8	89.3/84.6	90.7	87.7/90.9	70.4	92.5	88.7/88.4	81.46
6		1e-4	48.5	83.3/83.4	90.0/85.5	90.4	87.4/90.7	70.4	92.4	88.6/88.2	81.47
7		5e-4	16.2	74.9/76.4	89.7/84.8	87.8	85.0/89.0	68.2	91.6	86.2/86.5	75.01
8		Best (above)	48.5	83.3/83.8	90.0/85.5	90.7	87.7/90.9	70.4	92.5	88.7/88.4	81.59
9	Three-Stage	2e-5	49.3	83.1/83.6	89.5/84.3	91.0	87.6/90.9	70.8	92.4	88.7/88.3	81.57
10		1e-4	49.0	83.2/83.4	89.7/84.8	90.9	87.6/90.7	73.6	92.2	88.8/88.5	81.84
11		5e-4	29.5	81.3/81.9	89.1/83.6	88.4	85.0/89.3	66.1	91.9	84.1/83.9	77.34
12		Best (above)	49.3	83.2/83.4	89.7/84.8	91.0	87.6/90.9	73.6	92.4	88.8/88.5	81.93

3.3 The importance of data augmentation

Prior works augment the task-specific datasets using a text-editing technique by randomly replacing words in a sentence with their synonyms, based on their similarity measure on GloVe embeddings [18]. They then use the augmented dataset for task-specific compression of BERT models, observing improved accuracy for extreme quantization [56, 3]. They hypothesized that data augmentation (DA) is vital for compressing transformer models. However, as we have observed that the binarized networks were largely under-trained and had no clear advantage of multi-stage versus single-stage under a larger training budget, it raises the question of the necessity of DA.

To better understand the importance of DA for extreme compression, we compare the end-task performance of both 1-bit quantized $\text{BERT}_{\text{base}}$ models with and without data augmentation, based on previous findings in this paper that make extreme compression more robust. Table 4 shows results for the comparison of the 1-bit $\text{BERT}_{\text{base}}$ model under two different training budgets. Note that MNLI/QQP does not use DA; we repeat the results.

We find that regardless of whether shorter or longer iterations, removing DA leads to an average of 0.66 and 0.77 points drop in average GLUE score. Notably, the accuracy drops more than 1 point on smaller tasks such as CoLA, MPRC, RTE, STS-B. Furthermore, similar performance degradation is observed when removing DA from training an FP32 half-sized BERT model (more details will be introduced in § 4), regardless of whether using 1S-KD or 2S-KD. These results indicate that DA is helpful, especially for extreme compressed models; DA significantly benefits from learning more diverse/small data to mitigate accuracy drops.

> **Finding 3.** Training without DA hurts performance on downstream tasks for various compression tasks, especially on smaller tasks.

We remark that although our conclusion for DA is consistent with the finding in [3], we have performed a far more well-through experiment than [3] where they consider a single budget and 2S-KD only.

Table 4: **The Comparison between the results with and without data augmentation (DA).** Row 1-4 is for $\text{BERT}_{\text{base}}$ under 1-bit quantization using 1S-KD. Row 5-8 is for a layer-reduced $\text{BERT}_{\text{base}}$ (six-layer) under Budget-C without quantization (please see § 4 for more details).

#	Cost or Stages	Data Aug.	CoLA Mcc	MNLI-m/-mm Acc/Acc	MRPC F1/Acc	QNLI Acc	QQP F1/Acc	RTE Acc	SST-2 Acc	STS-B Pear/Spea	Avg. all	DIFF
1	Budget-A	✓	50.4	83.7/84.6	90.0/85.8	91.3	88.0/91.1	72.9	92.8	88.8/88.4	82.38	-0.77
2		✗	48.9	83.7/84.6	89.0/83.6	91.0	88.0/91.1	71.5	92.5	87.6/87.5	81.61	
3	Budget-C	✓	57.3	84.2/84.4	90.7/86.5	91.0	88.3/91.3	74.0	93.1	89.2/88.8	83.44	-0.66
4		✗	55.0	84.2/84.4	90.0/85.0	90.7	88.3/91.3	73.3	92.9	88.2/87.9	82.78	
5	One-stage	✓	56.9	84.5/84.9	89.7/84.8	91.7	88.5/91.5	71.8	93.5	89.8/89.4	83.27	-2.18
6		✗	52.2	84.5/84.9	88.4/82.4	91.4	88.5/91.5	63.9	92.9	86.5/86.3	81.09	
7	Two-stage	✓	56.0	84.1/84.2	89.7/84.8	91.5	88.2/91.3	71.8	93.3	89.4/89.1	82.93	-2.03
8		✗	50.2	84.1/84.2	88.7/83.1	90.8	88.2/91.3	64.6	92.9	86.9/86.7	80.90	

4 Interplay of KD, Long Training, Data Augmentation and Layer Reduction

In the previous section, we found strong evidence for statistically significant benefits from 1S-KD compared to 2S-KD and 3S-KD. Moreover, a considerable long training can greatly lift the score to the level same as the full precision teacher model. Although 1-bit $\text{BERT}_{\text{base}}$ already achieves $\times 32$ smaller size than its full-precision counterpart, a separate line of research focuses on changing the model architecture to reduce the model size. Notably, one of the promising methods – reducing model sizes via knowledge distillation – has long been studied and shown good performances [10, 55, 39, 49, 45].

While many works focus on innovating better knowledge distillation methods for higher compression ratio and better accuracy, given our observations in Section 3 that extreme quantization can effectively reduce the model size, we are interested in *to what extent can extreme quantization benefit state-of-the-art distilled models?*

To investigate this issue, we perform the following experiments: We prepare five student models, all having 6 layers with the same hidden dimension 768: (1) a pretrained TinyBERT_6 [18][1]; (2) MiniLMv2_6 [49] [2]; (3) Top-BERT_6: using the top 6 layers of the fine-tuned BERT-base model to initialize the student model; (4) Bottom-BERT_6: using the bottom six layers of the fine-tuned BERT-base model to initialize the student; (5) Skip-BERT_6: using every other layer of the fine-tuned BERT-based model to initialize the student. In all cases, we fine-tune $\text{BERT}_{\text{base}}$ for each task as the teacher model (see the teachers' performance in Table 5, Row 0). We choose TinyBERT and MiniLM because they are the state-of-the-art for knowledge distillation of BERT models. We choose the other three configurations because prior work [44, 32] also suggested that layer pruning is also a practical approach for task-specific compression. For the initialization of the student model, both TinyBERT_6 and MiniLM_6 are initialized with weights distilled from $\text{BERT}_{\text{base}}$ through pretraining distillation, and the other three students (e.g., top, bottom, skip) are obtained from the fine-tuned teacher model *without incurring any pretraining training cost*. We apply 1S-KD as we also verify

Table 5: Pre-training does not show benefits for layer reduction. Row 3 (rep.*) is a reproduced result by following the training recipe in [18].

#	Model	size	CoLA Mcc	MNLI-m/-mm Acc/Acc	MRPC F1/Acc	QNLI Acc	QQP F1/Acc	RTE Acc	SST-2 Acc	STS-B Pear/Spea	Avg. all	Acc. Drop
1	BERT-base fp32	417.2	59.7	84.9/85.6	90.6/86.3	92.1	88.6/91.5	72.2	93.2	90.1/89.6	83.95	-
Training cost: greater than Budget-C (see [18] or § B)												
2	Pretrained TinyBERT_6 ([18])	255.2 ($\times$1.6)	54.0	84.5/84.5	90.6/86.3	91.1	88.0/91.1	73.4	93.0	90.1/89.6	83.11	-0.84
3	Pretrained TinyBERT_6 (rep.*)	255.2 ($\times$1.6)	56.9	84.4/84.8	90.1/85.5	91.3	88.4/91.4	72.2	93.2	90.3/90.0	83.33	-0.62
Training cost: Budget-C												
4	Pretrained TinyBERT_6	255.2 ($\times$1.6)	54.4	84.6/84.3	90.4/86.3	91.5	88.5/91.5	69.7	93.3	89.2/89.0	82.76	-1.19
5	Pretrained MiniLM_6-v2	255.2 ($\times$1.6)	55.4	84.4/84.5	90.7/86.5	91.4	88.5/91.5	71.8	93.3	89.4/89.0	83.13	-0.82
6	Skip-BERT_6 (ours)	255.2 ($\times$1.6)	56.9	84.6/84.9	90.4/85.8	91.8	88.6/91.6	72.6	93.5	89.8/89.4	**83.50**	-0.45
7	Skip-BERT_5 (ours)	228.2 ($\times$1.8)	57.9	84.3/85.1	90.1/85.5	91.4	88.5/91.5	72.2	93.3	89.2/88.9	**83.38**	-0.57
8	Skip-BERT_4 (ours)	201.2 ($\times$2.1)	53.3	83.2/83.4	90.0/85.3	90.8	88.2/91.3	70.0	93.5	88.8/88.4	82.18	-1.77

that 2S-KD and 3S-KD are under-performed (See Table C.7 in the appendix). We set Budget-C as our training cost because the training budget in reproducing the results for TinyBert is greatly larger than Budget-C illustrated above. We report their best validation performances in Table 5 across the three learning rate {5e-5, 1e-4, 5e-4}. We apply 1S-KD and budget-C for this experiment. We report their

[1]The checkpoint of TinyBert is downloaded from their uploaded huggingface.co.

[2]The checkpoint of miniLMv2 is from their github.

best validation performances in Table 5 across three learning rates {5e-5, 1e-4, 5e-4}. For complete statistics with these learning rates, please refer to Table C.5. We make a few observations:

First, perhaps a bit surprisingly, the results in Table 5 show that there are no significant improvements from using the more computationally expensive pre-pretraining distillation in comparison to lightweight layer reduction: Skip-BERT$_6$ (Row 7) achieves the highest average score 83.50. This scheme achieves the highest score on larger and more robust tasks such as MNLI/QQP/QNLI/SST-2.

One noticeable fact in Table 5 is that under the same budget (Budget-C), the accuracy drop of Skip-BERT$_6$ is about 0.74 (0.37) higher than TinyBERT$_6$ in Row 5 (MINILM$_6$-v2 in Row 6). Note that layer reduction without pretraining has also been addressed in [38]. However, their KD is limited to logits without DA, and the performance is not better a trained DistillBERT. Second, with the above encouraging results from a half-size model (Skip-BERT$_6$), we squeeze the depths into five and four layers. The five-/four-layer student is initialized from ℓ-layer of teacher with $\ell \in \{3, 5, 7, 9, 11\}$ or $\ell \in \{3, 6, 9, 12\}$. We apply the same training recipe as Skip-BERT$_6$ and report the results in Row 8 and 9 in Table 5. There is little performance degradation in our five-layer model (83.28) compared to its six-layer counterpart (83.50); Interestingly, CoLA and MNLI-mm even achieve higher accuracy with smaller model sizes. Meanwhile, when we perform even more aggressive compression by reducing the depth to 4, the result is less positive as the average accuracy drop is around 1.3 points.

Third, among three lightweight layer reduction methods, we confirm that a Skip-# student performs better than those using Top-# and Bottom-#. We report the full results of this comparison in the Appendix Table C.6. This observation is consistent with [18]. However, their layerwise distillation method is not the same as ours, e.g, they use the non-adapted pretrained BERT model to initialize the student, whereas we use the fine-tuned BERT weights. We remark that our finding is in contrast to [38] which is perhaps because the KD in [38] only uses logits distillation without layerwise distillation.

Finding 4. Lightweight layer reduction matches or even exceeds expensive pre-training distillation for task-specific compression.

5 Proposed Method for Further Pushing the Limit of Extreme Compression

Based on our studies, we propose a simple yet effective method tailored for extreme lightweight compression. Figure 1 (right) illustrates our proposed method: XtrmC, which consists of only 2 steps: **Step I: Lightweight layer reduction.** Unlike the common layer reduction method where the layer-reduced model is obtained through computationally expensive pre-training distillation, we select a subset of the fine-tuned teacher weights as a lightweight layer reduction method (e.g., either through simple heuristics as described in Section 4 or search-based algorithm as described in [32]) to initialize the layer-reduced model. When together with the other training strategies identified in this paper, we find that such a lightweight scheme allows for achieving a much larger compression ratio while setting a new state-of-the-art result compared to other existing methods. **Step II: 1-bit quantization by applying 1S-KD with DA and long training.** Once we obtain the layer-reduced model, we apply the quantize-aware 1S-KD, proven to be the most effective in § 3.2. To be concrete, we use an ultra-low bit (1-bit/2-bit) quantizer to compress the layer-reduced model weights for a forward pass and then use STE during the backward pass for passing gradients. Meanwhile, we minimize the single-stage deep knowledge distillation objective with data augmentation enabled and longer training Budget-C (such that the training loss is close to zero).

Table 6: 1-/2-bit quantization of the layer-reduced model. The last column (Acc. drop) is the accuracy drop from their own fp-32 models. See full details in Table C.8.

#	Method	bit (#-layer)	size (MB)	CoLA Mcc	MNLI-m/-mm Acc/Acc	MRPC F1/Acc	QNLI Acc	QQP F1/Acc	RTE Acc	SST-2 Acc	STS-B Pear/Spea	Avg. all	Acc. drop
1	[56]	2 (6L)	16.0 (×26.2)	53.0	83.4/83.8	91.5/88.0	89.9	87.2/90.5	71.8	93.0	86.9/86.5	82.26	-0.76
2	Ours	2 (6L)	16.0 (×26.2)	53.8	83.6/84.2	90.5/86.3	90.6	88.2/91.3	73.6	93.6	89.0/88.7	**82.89**	-0.44
3		2 (5L)	14.2 (×29.3)	53.9	83.3/84.1	90.4/86.0	90.4	88.2/91.2	71.8	93.0	88.4/88.0	**82.46**	-0.72
4		2 (4L)	12.6 (×33.2)	50.3	82.5/83.0	90.0/85.3	89.2	87.8/91.0	69.0	92.8	87.9/87.4	81.22	-0.90
5		1 (6L)	8.0 (×52.3)	52.5	83.4/83.8	90.0/85.3	89.4	87.9/91.1	68.6	93.1	88.4/88.0	81.71	-1.51
6		1 (5L)	7.1 (×58.5)	52.2	82.9/83.2	89.9/85.0	88.5	87.6/90.8	69.3	92.9	87.3/87.0	81.34	-1.84
7		1 (4L)	6.3 (×66.4)	48.3	82.0/82.3	89.9/85.5	87.7	86.9/90.4	63.9	92.4	87.1/86.7	79.96	-2.01

Evaluation Results. The results of our approach are presented in Table 6, which we include multiple layer-reduced models with both 1-bit and 2-bit quantization. We make the following two observations:
(1) For the 2-bit + 6L model (Row 1 and Row 2), XtrmC achieves 0.63 points higher accuracy than 2-bit quantized model TernaryBERT [56]. This remarkable improvement consists of two factors: (a) our fp-32 layer-reduced model is better (b) when binarization, we train longer under three learning rate searches. To understand how much the factor-(b) benefits, we may refer to the accuracy drop from their fp32 counterparts (last col.) where ours only drops 0.44 points and TinyBERT$_6$ drops 0.76.
(2) When checking between our 2-bit five-layer (Row 3, 82.46) in Table 6 and 2-bit TinyBERT$_6$ (Row 1, 82.26), our quantization is much better, while our model size is 12.5% smaller, setting a new state-of-the-art result for 2-bit quantization with this size.
(3) Let us take a further look at accuracy drops (last column) for six/five/four layers, which are 0.44 (1.51), 0.72 (1.84), and 0.9 (2.01) respectively. It is clear that smaller models become much more brittle for extreme quantization than BERT$_{\text{base}}$, especially the 1-bit compression as the degradation is $\times 3$ or $\times 4$ higher than the 2-bit quantization.

Besides the reported results above, we have also done a well-thorough investigation on the effectiveness of adding a low-rank (LoRa) full-precision weight matrices to see if it will further push the accuracy for the 1-bit layer-reduced models. However, we found a negative conclusion on using LoRa. Interestingly, if we repeat another training with the 1-bit quantization after the single-stage long training, there is another 0.3 increase in the average GLUE score (see Table C.14). Finally, we also verify that prominent teachers can indeed help to improve the results (see appendix).

6 Conclusions

We carefully design and perform extensive experiments to investigate the contemporary existing extreme quantization methods [18, 3] by fine-tuning pre-trained BERT$_{\text{base}}$ models with various training budgets and learning rate search. Unlike [3], we find that there is no sharp accuracy drop if long training with data augmentation is used and that multi-stage KD and pre-training introduced in [18] is not a must in our setup. Based on the finding, we derive a user-friendly celebrating recipe for extreme quantization (see Figure 1, right), which allows us to achieve a larger compression ratio and higher accuracy. See our summarized results in Table 7.[3]

Table 7: Summary of task-specific performance of MNLI and GLUE scores. Also see Figure 2.

#	Model (8-bit activation)	size (MB)	MNLI-m/-mm	GLUE Score
1	BERT$_{\text{base}}$ (Teacher)	417.2 ($\times$1.0)	84.9/85.6	83.95
2	TinyBERT$_6$ (rep*)	255.2 ($\times$1.6)	84.4/84.8	83.33
3	XtrmC-BERT$_6$ (ours)	255.2 ($\times$1.6)	**84.6**/84.9	**83.50**
4	XtrmC-BERT$_5$ (ours)	228.2 ($\times$1.8)	84.3/**85.1**	**83.38**
5	2-bit BERT [56]	26.8 ($\times$16.0)	83.3/83.3	82.73
6	2-bit XtrmC-BERT (ours)	26.8 ($\times$16.0)	**84.6/84.7**	**84.10**
7	1-bit BERT (TWN) [3]	16.5 ($\times$25.3)	84.2/**84.7**	82.49
8	1-bit BERT (BWN) [3]	13.4 ($\times$32.0)	84.2/84.0	82.29
9	1-bit XtrmC-BERT (ours)	13.4 ($\times$32.0)	84.2/84.4	**83.44**
10	2-bit TinyBERT$_6$ [56]	16.0 ($\times$26.2)	83.4/83.8	82.26
11	2-bit XtrmC-BERT$_6$ (ours)	16.0 ($\times$26.2)	**83.6/84.2**	**82.89**
12	2-bit XtrmC-BERT$_5$ (ours)	14.2 ($\times$29.3)	83.3/84.1	**82.46**
13	1-bit XtrmC-BERT$_6$ (ours)	8.0 ($\times$52.3)	83.4/83.8	81.71
14	1-bit XtrmC-BERT$_5$ (ours)	7.1 ($\times$58.5)	82.9/83.2	81.34

Discussion and future work. Despite our encouraging results, we note that there is a caveat in our **Step I** when naively applying our method (i.e., initializing the student model as a subset of the teacher) to reduce the width of the model as the weight matrices dimension of teacher do not fit into that of students. We found that the accuracy drop is considerably sharp. Thus, in this domain, perhaps pertaining distillation [18] could be indeed useful, or Neural Architecture Search (AutoTinyBERT [53]) could be another promising direction. We acknowledge that, in order to make a fair comparison with [56, 3], our investigations are based on the classical 1-bit [36] and 2-bit [25] algorithms and without quantization on layernorm [2]. Exploring the benefits of long training on an augmented dataset with different 1-/2-bit algorithms (with or without quantization on layer norm) would be of potential interests [22, 33]. Finally, as our experiments focus on BERT$_{\text{base}}$ model, future work can be understanding how our conclusion transfers to decoder models such as GPT-2/3 [34, 6].

[3] We decide not to include many other great works [24, 54, 39, 33] since they do not use data augmentation or their setups are not close.

References

[1] Gustavo Aguilar, Yuan Ling, Yu Zhang, Benjamin Yao, Xing Fan, and Chenlei Guo. Knowledge distillation from internal representations. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, pages 7350–7357. AAAI Press, 2020.

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[3] Haoli Bai, Wei Zhang, Lu Hou, Lifeng Shang, Jin Jin, Xin Jiang, Qun Liu, Michael Lyu, and Irwin King. BinaryBERT: Pushing the limit of BERT quantization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4334–4348, Online, August 2021. Association for Computational Linguistics.

[4] Ron Banner, Itay Hubara, Elad Hoffer, and Daniel Soudry. Scalable methods for 8-bit training of neural networks. *Advances in neural information processing systems*, 31, 2018.

[5] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020.

[7] Tianlong Chen, Jonathan Frankle, Shiyu Chang, Sijia Liu, Yang Zhang, Zhangyang Wang, and Michael Carbin. The lottery ticket hypothesis for pre-trained bert networks. *Advances in neural information processing systems*, 33:15834–15846, 2020.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019)*, pages 4171–4186, 2019.

[9] Zhen Dong, Zhewei Yao, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Hawq: Hessian aware quantization of neural networks with mixed-precision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 293–302, 2019.

[10] Angela Fan, Edouard Grave, and Armand Joulin. Reducing transformer depth on demand with structured dropout. In *International Conference on Learning Representations*, 2020.

[11] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

[12] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*, 2020.

[13] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[14] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015.

[15] Lu Hou, Zhiqi Huang, Lifeng Shang, Xin Jiang, Xiao Chen, and Qun Liu. Dynabert: Dynamic bert with adaptive width and depth. *Advances in Neural Information Processing Systems*, 33:9782–9793, 2020.

[16] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4107–4115, 2016.

[17] Shankar Iyer, Nikhil Dandekar, and Kornl Csernai. First quora dataset release: Question pairs, 2017. *URL https://data. quora. com/First-Quora-Dataset-Release-Question-Pairs*, 2017.

[18] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online, November 2020. Association for Computational Linguistics.

[19] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is BERT really robust? A strong baseline for natural language attack on text classification and entailment. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 8018–8025. AAAI Press, 2020.

[20] Felix Juefei-Xu, Vishnu Naresh Boddeti, and Marios Savvides. Local binary convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 19–28, 2017.

[21] Minje Kim and Paris Smaragdis. Bitwise neural networks. *arXiv preprint arXiv:1601.06071*, 2016.

[22] Sehoon Kim, Amir Gholami, Zhewei Yao, Michael W Mahoney, and Kurt Keutzer. I-bert: Integer-only bert quantization. In *International conference on machine learning*, pages 5506–5518. PMLR, 2021.

[23] François Lagunas, Ella Charlaix, Victor Sanh, and Alexander M Rush. Block pruning for faster transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10619–10629, 2021.

[24] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[25] Fengfu Li, Bo Zhang, and Bin Liu. Ternary weight networks. *arXiv preprint arXiv:1605.04711*, 2016.

[26] 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. *CoRR*, abs/1907.11692, 2019.

[27] Zhenhua Liu, Yunhe Wang, Kai Han, Wei Zhang, Siwei Ma, and Wen Gao. Post-training quantization for vision transformer. *Advances in Neural Information Processing Systems*, 34, 2021.

[28] Yihuan Mao, Yujing Wang, Chufan Wu, Chen Zhang, Yang Wang, Quanlu Zhang, Yaming Yang, Yunhai Tong, and Jing Bai. Ladabert: Lightweight adaptation of BERT through hybrid model compression. In Donia Scott, Núria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 3225–3234. International Committee on Computational Linguistics, 2020.

[29] Naveen Mellempudi, Abhisek Kundu, Dheevatsa Mudigere, Dipankar Das, Bharat Kaul, and Pradeep Dubey. Ternary neural networks with fine-grained quantization. *arXiv preprint arXiv:1705.01462*, 2017.

[30] Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. On the stability of fine-tuning {bert}: Misconceptions, explanations, and strong baselines. In *International Conference on Learning Representations*, 2021.

[31] Markus Nagel, Rana Ali Amjad, Mart Van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or down? adaptive rounding for post-training quantization. In *International Conference on Machine Learning*, pages 7197–7206. PMLR, 2020.

[32] David Peer, Sebastian Stabinger, Stefan Engl, and Antonio Rodríguez-Sánchez. Greedy-layer pruning: Speeding up transformer models for natural language processing. *Pattern Recognition Letters*, 157:76–82, 2022.

[33] Haotong Qin, Yifu Ding, Mingyuan Zhang, Qinghua YAN, Aishan Liu, Qingqing Dang, Ziwei Liu, and Xianglong Liu. BiBERT: Accurate fully binarized BERT. In *International Conference on Learning Representations*, 2022.

[34] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[35] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683, 2019.

[36] Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. Xnor-net: Imagenet classification using binary convolutional neural networks. In *European conference on computer vision*, pages 525–542. Springer, 2016.

[37] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

[38] Hassan Sajjad, Fahim Dalvi, Nadir Durrani, and Preslav Nakov. Poor man's BERT: smaller and faster transformer models. *CoRR*, abs/2004.03844, 2020.

[39] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019.

[40] Victor Sanh, Thomas Wolf, and Alexander Rush. Movement pruning: Adaptive sparsity by fine-tuning. *Advances in Neural Information Processing Systems*, 33:20378–20389, 2020.

[41] Sheng Shen, Zhen Dong, Jiayu Ye, Linjian Ma, Zhewei Yao, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Q-bert: Hessian based ultra low precision quantization of bert. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8815–8821, 2020.

[42] Gil Shomron, Freddy Gabbay, Samer Kurzum, and Uri Weiser. Post-training sparsity-aware quantization. *Advances in Neural Information Processing Systems*, 34, 2021.

[43] Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv e-prints*, pages arXiv–2201, 2022.

[44] Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. Patient knowledge distillation for BERT model compression. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4322–4331, 2019.

[45] Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. Mobilebert: a compact task-agnostic bert for resource-limited devices. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2158–2170, 2020.

[46] Wei Tang, Gang Hua, and Liang Wang. How to train a compact binary neural network with high accuracy? In Satinder Singh and Shaul Markovitch, editors, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 2625–2631. AAAI Press, 2017.

[47] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations*, 2019.

[48] Lin Wang and Kuk-Jin Yoon. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *CoRR*, abs/2004.05937, 2020.

[49] Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. Minilmv2: Multi-head self-attention relation distillation for compressing pretrained transformers. *CoRR*, abs/2012.15828, 2020.

[50] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[51] Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*, 2017.

[52] Zhewei Yao, Linjian Ma, Sheng Shen, Kurt Keutzer, and Michael W Mahoney. Mlpruning: A multilevel structured pruning framework for transformer-based models. *arXiv preprint arXiv:2105.14636*, 2021.

[53] Yichun Yin, Cheng Chen, Lifeng Shang, Xin Jiang, Xiao Chen, and Qun Liu. Autotinybert: Automatic hyper-parameter optimization for efficient pre-trained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5146–5157, 2021.

[54] Ali Hadi Zadeh, Isak Edo, Omar Mohamed Awad, and Andreas Moshovos. Gobo: Quantizing attention-based nlp models for low latency and energy efficient inference. In *2020 53rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, pages 811–824. IEEE, 2020.

[55] Minjia Zhang and Yuxiong He. Accelerating training of transformer-based language models with progressive layer dropping. *Advances in Neural Information Processing Systems*, 33:14011–14023, 2020.

[56] Wei Zhang, Lu Hou, Yichun Yin, Lifeng Shang, Xiao Chen, Xin Jiang, and Qun Liu. Ternary-BERT: Distillation-aware ultra-low bit BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 509–521, Online, November 2020. Association for Computational Linguistics.

[57] Sanqiang Zhao, Raghav Gupta, Yang Song, and Denny Zhou. Extremely small BERT models from mixed-vocabulary training. In Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 2753–2759, 2021.

[58] Minjia Zhng, Niranjan Uma Naresh, and Yuxiong He. Adversarial data augmentation for task-specific knowledge distillation of pre-trained transformers. *Association for the Advancement of Artificial Intelligence*, 2022.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Please see § 6.

(c) Did you discuss any potential negative societal impacts of your work? [N/A] XtrmC a common machine learning technique.

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...

(a) Did you state the full set of assumptions of all theoretical results? [N/A]

(b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See details in Appendix B.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] All details are in main text as well as Appendix B.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A] Our results are based on a single random seed for all setup. So it does not rely on the randomness. However, we do provide the error bars for a set of experiments to verify even with randomness, the error by changing random seed does not at all affect our conclusion in Appendix C.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Those are in main text.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes] We cited all dataset and github repository used in the paper.

(b) Did you mention the license of the assets? [N/A]

(c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]