# SKDBERT: Compressing BERT via Stochastic Knowledge Distillation

**Zixiang Ding**[*1], **Guoqing Jiang**[1], **Shuai Zhang**[1], **Lin Guo**[1], **Wei Lin**[2]

[1]Meituan
[2]Individual
{dingzixiang, jiangguoqing03, zhangshuai51, guolin08}@meituan.com, lwsaviola@163.com

## Abstract

In this paper, we propose Stochastic Knowledge Distillation (SKD) to obtain compact BERT-style language model dubbed SKDBERT. In each iteration, SKD samples a teacher from a pre-defined teacher ensemble, which consists of multiple teachers with multi-level capacities, to transfer knowledge into student in an one-to-one manner. Sampling distribution plays an important role in SKD. We heuristically present three types of sampling distributions to assign appropriate probabilities for multi-level teachers. SKD has two advantages: 1) it can preserve the diversities of multi-level teachers via stochastically sampling single teacher in each iteration, and 2) it can also improve the efficacy of knowledge distillation via multi-level teachers when large capacity gap exists between the teacher and the student. Experimental results on GLUE benchmark show that SKDBERT reduces the size of a BERT model by 40% while retaining 99.5% performances of language understanding and being 100% faster.

## Introduction

BERT-style (Devlin et al. 2019) language models, e.g., XL-Net (Yang et al. 2019), RoBERTa (Liu et al. 2019), T5 (Raffel et al. 2020), ELECTRA (Clark et al. 2020), have achieved amazing performance in natural language processing. However, the numerous parameters of the above BERT-style language models greatly increase the difficulty of deployment on resource-constrained devices. Recently, some works have demonstrated that many parameters are redundant in BERT-style language models (Michel, Levy, and Neubig 2019; Voita et al. 2019; Kovaleva et al. 2019). For instance, Voita et al. (2019) claim that reducing the head number of BERT does not result in quality degradation. Consequently, many compression approaches have been proposed to obtain resource-friendly BERT-style language models, e.g., parameter sharing based (Lan et al. 2020), Knowledge Distillation (KD) based (Iandola et al. 2020; Xu et al. 2020; Pan et al. 2021), pruning based (Fan, Grave, and Joulin 2020; Guo et al. 2019), quantization based (Shen et al. 2020) and NAS-based (Chen et al. 2020; Xu et al. 2021, 2022). In this paper, we focus on the KD-based approaches.

[*]Corresponding author


| Teacher | MRPC $\frac{\text{F1+Acc}}{2}$ | RTE Acc | SST-2 Acc | QQP $\frac{\text{F1+Acc}}{2}$ | QNLI Acc | MNLI m |
|---|---|---|---|---|---|---|
| $T_0$:8-768-12 | 89.6 | 73.3 | 92.0 | 88.9 | 91.1 | 82.9 |
| $T_1$:10-768-12 | 89.7 | 71.8 | 92.3 | **89.0** | 91.3 | 83.2 |
| $T_2$:12-768-12 | 89.1 | 71.5 | **93.1** | 88.9 | **91.4** | 82.8 |
| $T_3$:24-1024-16 | **90.0** | 72.9 | 92.1 | 88.9 | 91.2 | 83.4 |
| $T_4$:24-1024-16† | 89.5 | 72.6 | 92.4 | **89.0** | 91.3 | 83.5 |
| $T_0$-$T_4$ | 89.7 | **73.7** | 92.2 | 88.6 | 91.1 | **83.6** |

Table 1: Distillation performances of our student with single and multiple teachers on the development set of GLUE benchmark (Wang et al. 2019). Moreover, $T_i$:$r$-$s$-$t$ indicates that the layer number, hidden size and head number of $i$-th teacher are $r$, $s$ and $t$, respectively. † means the teacher is pre-trained with whole word masking.

The main differences among the KD-based BERT-style language model compression approaches are:

- Learning procedure: pre-training-only (Turc et al. 2019; Sanh et al. 2019; Sun et al. 2020), fine-tuning-only (Sun et al. 2019; Wu, Wu, and Huang 2021), and both pre-training and fine-tuning (Jiao et al. 2020).
- Distillation objective: soft target probabilities (Sanh et al. 2019; Sun et al. 2020; Wu, Wu, and Huang 2021), embedding outputs (Jiao et al. 2020), hidden states (Sun et al. 2020; Jiao et al. 2020; Wu, Wu, and Huang 2021), self-attention distributions (Sun et al. 2020; Jiao et al. 2020; Wang et al. 2020) and self-attention value relation (Wang et al. 2020).

Wu, Wu, and Huang (2021) employ multiple teachers to achieve better performance than single-teacher KD based approaches on several downstream tasks. However, we find that the ensemble of multiple teachers can not always outperform single teacher for knowledge distillation, as shown in Table 1[1] Two possible reasons for the above phenomenon are: 1) the ensemble prediction of the teachers loses diversity (Tran et al. 2020), and 2) the large capacity gap between the teacher and the student impacts the efficacy of knowledge distillation (Mirzadeh et al. 2020).

To solve the above mentioned issues, we propose Stochas-

[1]The implementation details can be found in Section A of supplementary materials at https://arxiv.org/pdf/2211.14466.pdf.

tic Knowledge Distillation (SKD) to obtain compact BERT-style language model dubbed SKDBERT. SKD focuses on distillation paradigm rather than the learning procedure and the distillation objective. In each iteration, SKD samples a teacher from a pre-defined teacher ensemble, which consists of multiple teachers with multi-level capacities, to transfer knowledge into student in an one-to-one manner. Sampling distribution plays an important role in SKD. We heuristically present three types of sampling distributions to assign appropriate sampling probability for each teacher. The proposed SKD is effective to solve the above issues:

- For the issue of losing diversity: In each iteration, SKD directly learns knowledge from a sampled teacher in the one-to-one manner to preserve the diversity of each teacher as much as possible.
- For the issue of capacity gap: In entire distillation procedure, SKD can not only utilize weak teachers to fill the capacity gap between the strong teacher and the student, but also avoid limiting the performance of the student to the weak teachers.

To examine the generalization ability of SKD, we have also conducted image classification experiments with ResNet (He et al. 2016) and its variants (Zagoruyko and Komodakis 2016) on CIFAR-100 (see Section B of supplementary materials in detail).

In summary, our contribution is two-fold: 1) We propose Stochastic Knowledge Distillation (SKD) to obtain compact BERT-style language model dubbed SKDBERT, and show its superiority for the issues of losing diversity and capacity gap. 2) Extensive experiments on the GLUE benchmark show that SKDBERT reduces the size of $\text{BERT}_{\text{BASE}}$ by 40% while retaining its 99.5% performances of language understanding and being 100% faster.

## Related Work

**Knowledge Distillation** KD consists of three components: 1) knowledge type, e.g., response-based (Hinton, Vinyals, and Dean 2015), feature-based (Yang et al. 2021), relation-based (Yang et al. 2022), 2) teacher-student architecture, e.g., simplified (Li et al. 2020), quantized (Polino, Pascanu, and Alistarh 2018), condensed (Xie et al. 2020), and 3) distillation strategy, e.g., multi-teacher (Yuan et al. 2021), graph-based (Yao et al. 2020), adversarial (Micaelli and Storkey 2019). The proposed SKD relates to multi-teacher distillation algorithms, especially Weighted-response (WKD) (Wu, Wu, and Huang 2021) and Teacher-Assistant (TAKD) (Mirzadeh et al. 2020) which are introduced in detail in next section.

**Knowledge Distillation for BERT** BERT (Devlin et al. 2019) has been compressed by various KD-based approaches. DistilBERT (Sanh et al. 2019), MobileBERT (Sun et al. 2020) and MiniLM (Wang et al. 2020) adopt different strategies for distillation in the pre-training stage. MobileBERT transfers knowledge progressively from a specific inverted-bottleneck BERT. MiniLM proposes deep self-attention distillation, which aims to minimize the KL-divergence between the value relation of the teacher and the student. In the fine-tuning stage, BERT-PKD (Sun et al. 2019) uses the combination of response-based and feature-based knowledge for BERT-style language model compression. Furthermore, TinyBERT (Jiao et al. 2020) proposes a two-state distillation framework, which learns response-based and feature-based knowledge simultaneously in both pre-training and fine-tuning phases. Moreover, data augmentation technique is also used to further improve the performance of TinyBERT. Besides, multi-teacher KD has also been used for BERT compression (Wu, Wu, and Huang 2021).

## The Proposed Approach

### Stochastic Knowledge Distillation

**Overview** The overview of SKD is shown in Figure 1 (c). In each iteration, SKD selects a teacher from a pre-defined teacher ensemble with a specific probability distribution $\pi(\cdot)$. In particular, the teacher ensemble consists of multiple fine-tuned teachers with multi-level capacities on specific downstream tasks. From a local perspective, SKD directly learns knowledge from the sampled teacher in the one-to-one manner to preserve the diversity. From a global perspective, SKD utilizes weak teachers to fill the large capacity gap between the teacher and student while avoiding limiting the performance of the student to the weak teachers.

There are two important components in SKD, i.e., the teacher ensemble and the sampling distribution. We discuss the influence of different teacher ensembles in terms of capacity and quantity in ablation studies. Sampling distribution $\pi(\cdot)$ determines the occurrence frequency of each teacher in entire distillation procedure. For various teacher ensembles and downstream tasks, the most appropriate distribution may be different. Consequently, we propose uniform, teacher-rank and student-rank sampling distributions for various cases.

**Uniform Distribution** means that the probability of each teacher is equal, and can be expressed as

$$\pi(n) = [p_1 = \frac{1}{n}, \cdots, p_n = \frac{1}{n}], \tag{1}$$

where $n$ is the number of teachers in the teacher ensemble, and $p_i (i = 1, \cdots, n)$ represents the sampling probability of $i$-th teacher.

**Teacher-rank Distribution** determines the probability in the light of fine-tuning performance $acc_{ft}$ of each teacher, and can be obtained by

$$\pi(n, acc_{ft}) = [p_1 = \frac{s_1^{ft}}{\sum_{j=1}^{n} j}, \cdots, p_n = \frac{s_n^{ft}}{\sum_{j=1}^{n} j}], \tag{2}$$

where $s_i^{ft} = n - r_i^{ft} + 1 \in [1, \cdots, n] (i = 1, 2, \cdots, n)$ refers to the fine-tuning performance score of $i$-th teacher with respect to $r_i^{ft} \in [1, \cdots, n]$ (the rank of $acc_{ft}$ which can be found in Section F of supplementary materials).

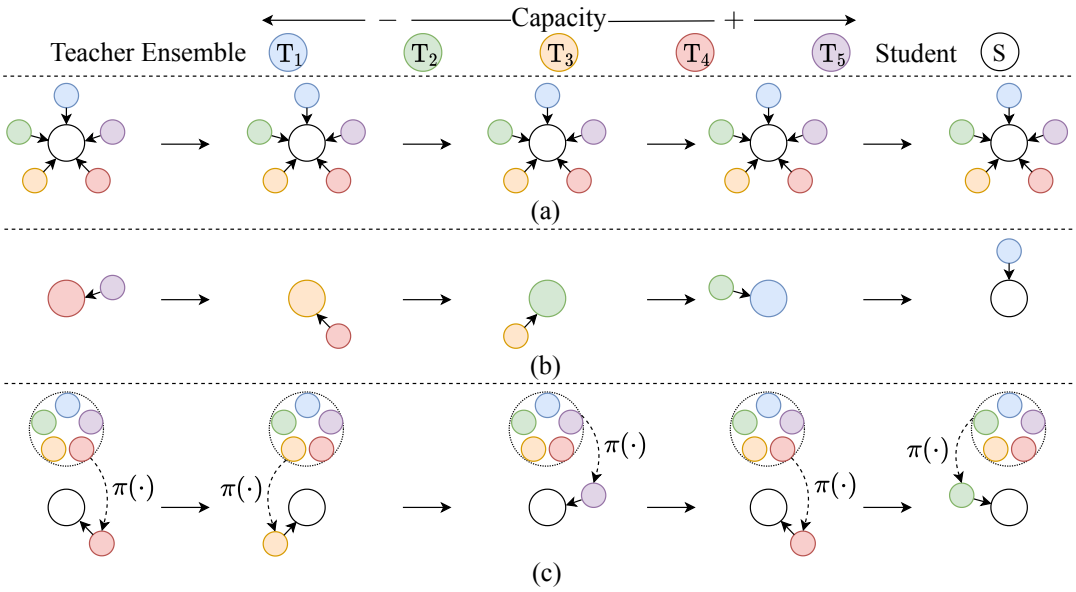



Figure 1: Comparison of WKD, TAKD and SKD. (a) WKD: The weighted logits with respect to all teachers in the teacher ensemble is used to optimize the student S in entire training procedure. (b) TAKD: Each teacher (except the strongest one, i.e., $T_5$) is progressively distilled by its predecessor according to the order of capacities. Subsequently, the weakest teacher is used for the student distillation. (c) SKD: According to a specific probability distribution $\pi(\cdot)$, in each iteration, a teacher is stochastically sampled from the teacher ensemble to distill the student in the one-to-one manner. Best viewed in color.

**Student-rank Distribution** is based on the distillation performance $acc_{dis}$ of the student with regard to each teacher. It contributes to filling the possible capacity gap between the teacher and the student, and can be calculated by

$$\pi(n, acc_{dis}) = [p_1 = \frac{s_1^{dis}}{\sum_{j=1}^{n} j}, \cdots, p_n = \frac{s_n^{dis}}{\sum_{j=1}^{n} j}], \quad (3)$$

where $s_i^{dis} = n - r_i^{dis} + 1 \in [1, \cdots, n](i = 1, 2, \cdots, n)$ refers to the distillation performance score of student distilled by $i$-th teacher with respect to $r_i^{dis} \in [1, \cdots, n]$ (the rank of $acc_{dis}$ which can be found in Section G of supplementary materials).

**SKD Learning** In this paper, SKD focuses on vanilla response-based distillation paradigm, i.e., only using logits as the knowledge to be transferred between the prediction layers of the teacher and the student. Given a specific $\pi(\cdot)$, let $f_{T\sim\pi(\cdot)}$ and $f_S$ denote the logits from the teacher and the student, respectively. Furthermore, the objective function of SKD can be expressed as

$$\mathcal{L}_{\text{SKD}} = \sum_{x \in \mathcal{X}} \mathcal{L}_d(f_{T\sim\pi(\cdot)}(x), f_S(x)),$$

where $\mathcal{L}_d$ represents distilled loss function, $\mathcal{X}$ refers to all training data. We discuss why does SKD works in Section C of supplementary materials.

## Comparison Against WKD and TAKD

The comparison of WKD, TAKD and SKD is shown in Figure 1. In entire distillation procedure, WKD employs the ensemble logits of $n = 5$ teachers as the knowledge, which loses the diversity. For TAKD, except the strongest teacher (i.e., $T_5$), each teacher is progressively distilled by its predecessor according to the order of capacities of teachers in each stage. Subsequently, the student is distilled by the weakest teacher (i.e., $T_1$) whose capacity plays an important role for the distillation performance. In each iteration, SKD stochastically samples a teacher for the student with a specific $\pi(\cdot)$ in the one-to-one distillation manner.

Different from WKD, SKD does not only learn abundant knowledge from the teacher ensemble, but also preserve diversity from each distinctive teacher. Compared to TAKD, SKD can employ multiple teachers with multi-level capacities to fill the possible capacity gap issue between the teacher and the student, while avoiding the drawback of sensitiveness for weak-capacity teachers. Furthermore, we show the comparison among WKD, TAKD and SKD on the GLUE benchmark in ablation studies.

# Experiments

## Datasets

We evaluate SKDBERT on the GLUE benchmark including MRPC (Dolan and Brockett 2005), RTE (Bentivogli et al. 2009), STS-B (Cer et al. 2017), SST-2 (Socher et al. 2013), QQP (Chen et al. 2018), QNLI (Rajpurkar et al. 2016) and MNLI (Williams, Nangia, and Bowman 2017).

## SKDBERT Settings

We employ the proposed SKD to obtain a 4-layer and a 6-layer BERT-style language models dubbed $\text{SKDBERT}_4$ and $\text{SKDBERT}_6$, respectively. More architecture details can be

| Model | #Params (M) | #FLOPs (B) | Speedup | MRPC | RTE | STS-B | SST-2 | QQP | QNLI | MNLI-m | MNLI-mm | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\text{BERT}_{\text{BASE}}$ (Devlin et al. 2019) | 109 | 22.5 | 1.0 × | 88.9 | 66.4 | 85.8 | 93.5 | 71.2 | 90.5 | 84.6 | 83.4 | 83.0 |
| $\text{BERT}_{\text{TINY}}$ (Devlin et al. 2019)† | 14.5 | 1.2 | 9.4 × | 83.2 | **62.6** | **77.1** | 87.6 | 66.5 | 84.8 | 75.4 | 74.9 | 76.5 |
| $\text{BERT}_{\text{SMALL}}$ (Devlin et al. 2019)† | 29.2 | 3.4 | 5.7 × | 83.4 | 61.8 | 77.0 | 89.7 | 68.1 | **86.4** | 77.6 | 77.0 | 77.0 |
| $\text{DistilBERT}_4$ (Sanh et al. 2019)† | 52.2 | 7.6 | 3.0 × | 82.4 | 54.1 | 76.1 | **91.4** | **68.5** | 85.2 | 78.9 | 78.0 | 76.8 |
| $\text{SKDBERT}_4$ (Ours) | 14.5 | 1.2 | 9.4 × | **85.8** | 62.4 | 76.9 | 88.1 | 68.0 | 85.1 | **79.3** | **78.3** | **78.1** |
| $\text{BERT}_6$-PKD (Sun et al. 2019)‡ | 66.0 | 11.3 | 2.0 × | 85.0 | 65.5 | 81.6 | 92.0 | 70.7 | 89.0 | 81.5 | 81.0 | 80.8 |
| PD (Turc et al. 2019) | 66.0 | 11.3 | 2.0 × | 86.8 | 65.3 | - | 91.8 | 70.4 | 88.9 | 82.8 | 82.2 | - |
| $\text{DistilBERT}_6$(Sanh et al. 2019)† | 66.0 | 11.3 | 2.0 × | 86.9 | 58.4 | 81.3 | **92.5** | 70.1 | 88.9 | 82.6 | 81.3 | 80.3 |
| BERT-of-Theseus (Xu et al. 2020) | 66.0 | 11.3 | 2.0 × | 87.6 | 66.2 | **84.1** | 92.2 | **71.6** | 89.6 | 82.4 | 82.1 | 82.0 |
| $\text{SKDBERT}_6$ (Ours) | 66.0 | 11.3 | 2.0 × | **88.4** | **68.8** | 83.9 | 91.5 | 71.4 | **90.3** | **83.4** | **82.8** | **82.6** |

Table 2: Results of SKDBERT and other popular approaches on GLUE-test. The inference speedup is obtained on a single NVIDIA K80 GPU. † and ‡ indicate that these results are cited from Jiao et al. (2020) and Xu et al. (2020), respectively. MRPC and QQP tasks are evaluated by F1 score, STS-B task is evaluated by Spearman correlations, and other tasks are evaluated by accuracy score.

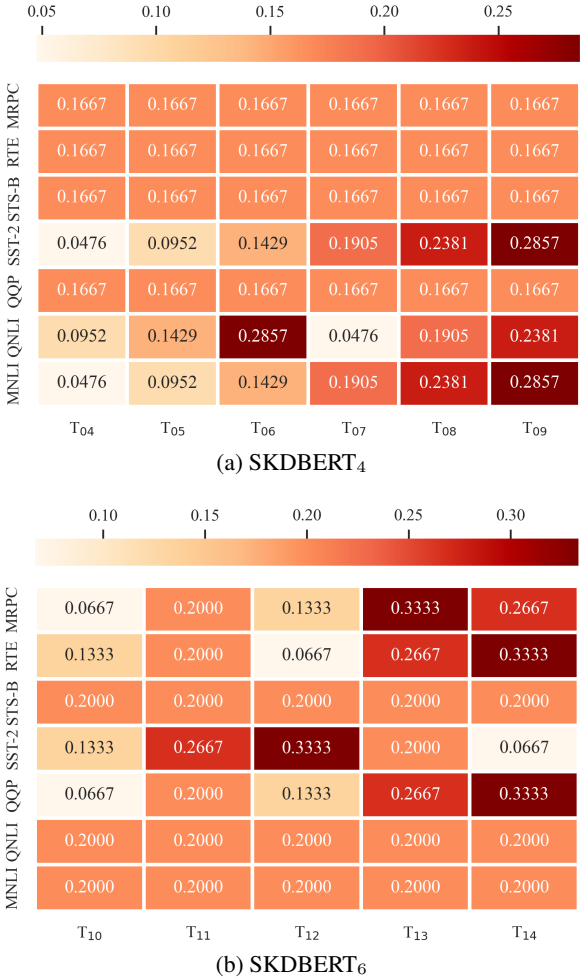
found in Section F of supplementary materials. Detailed introduction about the GLUE benchmark is given in Section D of supplementary materials. Unless otherwise specified, we employ the evaluation metrics shown in Section E of supplementary materials to select the best-performing model for different downstream tasks.

## Results on GLUE Benchmark

**Implementation Details** In this section, the teacher ensembles of $\text{SKDBERT}_4$ an $\text{SKDBERT}_6$ consist of 6 and 5 teachers, i.e., $\text{T}_{04}$-$\text{T}_{09}$ and $\text{T}_{10}$-$\text{T}_{14}$ whose architecture information can be found in Section F of supplementary materials, respectively. Moreover, fine-tuning performance of each teacher and distillation performance of SKDBERT with regard to each teacher can be found in Section F and G of supplementary materials, respectively. Subsequently, three types of sampling probabilities of each teacher ensemble can be derived from (1) to (3). According to the hyperparameters shown in Section E of supplementary materials, SKDBERT with the best performance on the development set of the GLUE benchmark (GLUE-dev) is submitted to the official GLUE evaluation server[2] to obtain results on the test set of GLUE benchmark (GLUE-test). Moreover, all implementations are performed on NVIDIA A100 GPU.

**Learned Distributions** For different downstream tasks, we employ various sampling distributions for $\text{SKDBERT}_4$ and $\text{SKDBERT}_6$ as shown in Figure 2.

**Results and Analysis** Table 2 summarizes the performance of SKDBERT and other popular compact BERT-style language models on GLUE-test. Both $\text{SKDBERT}_4$ and $\text{SKDBERT}_6$ achieve the best GLUE score compared to the prior state of the art. Furthermore, $\text{SKDBERT}_6$ outperforms all comparative approaches on the GLUE benchmark except STS-B, SST-2 and QQP tasks. $\text{SKDBERT}_6$ obtains 82.6% GLUE score which are only 0.4% less than $\text{BERT}_{\text{BASE}}$, using 66M parameters which are 38.5% less than $\text{BERT}_{\text{BASE}}$ with 109M parameters. Moreover, $\text{SKDBERT}_6$ outperforms $\text{BERT}_{\text{BASE}}$ on RTE and QQP tasks. Besides, TinyBERT (Jiao et al. 2020) achieves novel performance on the GLUE benchmark via extra learning procedure and distillation objective. We show a fair comparison between SKDBERT and TinyBERT in ablation studies.



(a) $\text{SKDBERT}_4$

(b) $\text{SKDBERT}_6$

Figure 2: The used distributions for SKDBERT.

[2]https://gluebenchmark.com

| Model | MRPC F1/Acc | RTE Acc | STS-B Pear/Spea | SST-2 Acc | QQP F1/Acc | QNLI Acc | MNLI-m Acc | MNLI-mm Acc |
|---|---|---|---|---|---|---|---|---|
| Poor Man's $BERT_6$ (Sajjad et al. 2020) | -/80.2 | 65.0 | -/88.5 | 90.3 | -/90.4 | 87.6 | 81.1 | - |
| LayerDrop (Fan, Grave, and Joulin 2020) | 85.9/- | 65.2 | -/85.7 | 90.7 | -/88.3 | 88.4 | 80.7 | - |
| BERT-PKD (Sun et al. 2019) | 85.7/- | 66.5 | -/86.2 | 91.3 | -/88.4 | 88.4 | 81.3 | - |
| BERT-of-Theseus (Xu et al. 2020) | **89.0**/- | 68.2 | -/**88.7** | 91.5 | -/89.6 | 89.5 | 82.3 | - |
| $MiniLM_6$ (Wang et al. 2020) | 88.4/- | 71.5 | -/- | 92.0 | -/**91.0** | 91.0 | 84.0 | - |
| $SKDBERT_6$ (Ours) | **89.0/92.1** | **75.5** | **89.2/88.7** | **92.9** | **87.9/91.0** | **91.4** | **84.1** | **83.7** |

Table 3: Results of SKDBERT and other popular approaches on GLUE-dev. All comparative approaches have identical architecture, i.e., 6-layer BERT-style language model with 66M parameters.

**More Comparisons** We compare SKDBERT to more compact BERT-style language models (e.g., Poor Man's BERT (Sajjad et al. 2020), BERT-of-Theseus (Xu et al. 2020) and MiniLM (Wang et al. 2020)) on GLUE-dev, and show the results in Table 3. Moreover, all the comparative models have identical architecture with $SKDBERT_6$. The proposed SKDBERT achieves the best performance on all tasks. In particular, on the task of RTE, SKDBERT achieves 4% accuracy score improvement over MiniLM which ranks the best in the comparative methods.

## Ablation Studies

We perform extensive ablation experiments to show the effectiveness of SKDBERT in terms of teacher ensemble, sampling distribution, KD paradigm, extra learning procedure and distillation objective. Appropriately increasing the number of teachers can effectively improve the diversity of prediction (Allingham et al. 2021) for obtaining better performance. As a result, we discuss the effectiveness of weak teachers (e.g., $T_{01}$ to $T_{03}$ for $SKDBERT_4$, $T_{01}$ to $T_{06}$ for $SKDBERT_6$). Moreover, the sampling distributions used in this section can be found in Section H of supplementary materials.

### Impact of Teacher Ensemble

**Capacity** We perform a list of experiments for SKDBERT with various teacher ensembles—$T_{04}$ to $T_{06}$, $T_{07}$ to $T_{09}$, $T_{10}$ to $T_{12}$ and $T_{12}$ to $T_{14}$—on GLUE-dev, to verify the influence of teacher ensemble's capacity, and show the results in Figure 3.

For $SKDBERT_4$, its performance is not proportionate to the increased teacher ensemble's capacity. Similar to vanilla KD, SKD suffers from the capacity gap issue (Mirzadeh et al. 2020), but which is greatly alleviated as shown in Figure 3, where the teacher ensemble of $T_{12}$ to $T_{14}$ achieves almost identical performance to the teacher ensemble of $T_{07}$ to $T_{09}$[3]. The performance of $SKDBERT_6$ is proportionate to the increased teacher ensemble's capacity. For $SKDBERT_6$ with vanilla KD, the performance of $T_{14}$ is almost identical to the weaker $T_{09}$ due to the capacity gap issue. However, $SKDBERT_6$ with the teacher ensemble of $T_{12}$ to $T_{14}$ achieves 0.75% higher GLUE score than the teacher ensemble of $T_{07}$ to $T_{09}$, due to the capacity gap alleviation ability of SKD. SKDBERT with the teacher ensemble of $T_{04}$ to $T_{06}$ whose capacities are all weaker than the student, achieves 1.35% higher GLUE score than vanilla KD with $T_{06}$ (about 84.9%) via the diversity extracted from $T_{04}$ and $T_{05}$. In particular, $SKDBERT_6$ without KD achieves 85.9% GLUE score which is 1% better than the one distilled by vanilla KD with $T_{06}$. Consequently, the proposed SKD can obtain knowledge from the teachers, whose capacities are weaker than the student.

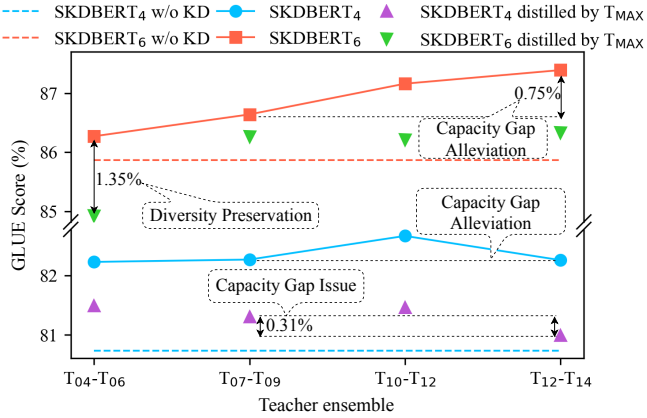



Figure 3: Impact of different-capacity teacher ensembles for SKDBERT on GLUE-dev. $T_{MAX}$ means the strongest teacher in the teacher ensemble, e.g., the $T_{MAX}$ is $T_{12}$ in the teacher ensemble of $T_{10}$ to $T_{12}$.

Above all, the teacher ensemble should contain strong-capacity teachers while taking the capacity gap issue into consideration to achieve novel performance for SKDBERT. Moreover, teachers whose capacities are weaker than the student, may contribute to improving the distillation performance via diversity preservation.

**Quantity** It is difficult to guarantee the invariable capacity of teachers when the number of teachers changes. Hence, we use two cases which have identical junior ($T_{04}$ to $T_{06}$ in case 1) and senior teachers ($T_{12}$ to $T_{14}$ in case 2) with respect to various teacher ensembles, to discuss the influence of teachers' quantity on the teacher ensemble, and show the results in Figure 4, where the results of $SKDBERT_6$ with various teacher ensembles shown in Figure 3 are given again for comprehensive comparison. Moreover, we fit the GLUE score of case 2 to obtain the curve with respect to $SKDBERT_6$.

[3]As the teacher for $SKDBERT_4$, $T_{14}$ performs 0.31% worse than $T_{09}$ due to the capacity gap issue.

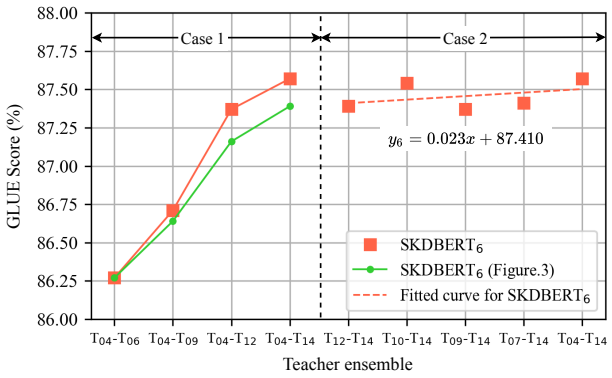



Figure 4: Impact of different-quantity teacher ensembles for SKDBERT$_6$ on GLUE-dev. In case 1, four groups of teacher ensembles have identical junior teachers, i.e., T$_{04}$ to T$_{06}$. In case 2, five groups of teacher ensembles have identical senior teachers, i.e., T$_{12}$ to T$_{14}$.

In case 1, SKDBERT$_6$ is proportionate to the increased teachers' quantity that is similar to the tendency of SKDBERT$_6$ shown in Figure 3 with little performance improvements. We consider that the primary reason of this phenomenon is the increased teacher ensemble's capacity. Besides, increasing number of teachers contributes to improving the performance of SKDBERT, but the improvement is limited[4] when the strongest teacher has powerful capacity.

In case 2, SKDBERT$_6$ is not sensitive to the quantity of teachers when various teacher ensembles contain identical senior teachers. We find that the slope of the fitted curve is only 0.023 for SKDBERT$_6$. Consequently, we indicate that the quantity of teachers hasn't distinct effect on performance improvement of SKDBERT that reaches the same conclusion with case 1.

Above all, we conclude that 1) increasing the quantity of junior teachers may contribute to improving the performance, but the improvement is limited, and 2) increasing the quantity of senior teachers is able to dramatically improve the performance.

## Impact of Sampling Distribution

In this section, we conduct similar-capacity case 1, which contains four teacher ensembles of T$_{04}$-T$_{06}$, T$_{07}$-T$_{09}$, T$_{10}$-T$_{12}$, T$_{12}$-T$_{14}$, and large-capacity-gap case 2, which consists of four teacher ensembles of T$_{04}$-T$_{06}$, T$_{04}$-T$_{09}$, T$_{04}$-T$_{12}$, T$_{04}$-T$_{14}$ for SKD, and show the results in Table 4.

Uniform distribution shows satisfactory performance with the similar-capacity teacher ensemble instead of with the large-capacity-gap one. In case 1, uniform distribution achieves the best performance with the teacher ensembles of T$_{07}$ to T$_{09}$, T$_{10}$ to T$_{12}$ and T$_{12}$ to T$_{14}$. However, in case 2, uniform distribution ranks the worst for three out of four teacher ensembles. Hence, junior teachers can only improve the performance of SKDBERT with low sampling probability. Compared to case 1, teacher-rank distribution performs better in case 2. Teacher-rank distribution obtains the worst performance for all teacher ensembles in case 1. Teacher-rank distribution performs better than uniform distribution in case 2. Student-based distribution delivers novel performance for both cases.

Furthermore, we use standard deviation as the metric to evaluate the performance difference among three types of distribution. In similar-capacity case 1, three sampling distributions have less difference. Nevertheless, there has an apparent difference among three distributions in large-capacity-gap case 2. The standard deviations are incremental from 0.08 to 0.26 with the capacity gap between the weakest and strongest teachers in the teacher ensemble.

Above all, uniform distribution is appropriate for similar-capacity teacher ensemble where the knowledge of each teacher contributes to improving the performance of SKDBERT. Furthermore, teacher-rank and student-rank distributions are appropriate for large-capacity-gap teacher ensemble, where the knowledge provided by junior teacher is prone to delivering negative effect for SKDBERT.

| Case | Ensemble | GLUE Score (%) | | | $\sigma$(%) |
|---|---|---|---|---|---|
| | | Uniform | Teacher-rank | Student-rank | |
| 1 | T$_{04}$-T$_{06}$ | 86.20 | 86.07 | **86.27** | 0.08 |
| | T$_{07}$-T$_{09}$ | **86.57** | 86.54 | **86.57** | 0.01 |
| | T$_{10}$-T$_{12}$ | **87.09** | 86.93 | 87.00 | 0.07 |
| | T$_{12}$-T$_{14}$ | **87.27** | 87.00 | 87.11 | 0.11 |
| 2 | T$_{04}$-T$_{06}$ | 86.20 | 86.07 | **86.27** | 0.08 |
| | T$_{04}$-T$_{09}$ | 86.33 | 86.57 | **86.63** | 0.13 |
| | T$_{04}$-T$_{12}$ | 86.77 | 86.89 | **87.33** | 0.24 |
| | T$_{04}$-T$_{14}$ | 86.84 | 87.10 | **87.47** | 0.26 |

Table 4: Impact of sampling distributions for SKDBERT$_6$ on GLUE-dev. For each teacher ensemble in case 1, teachers have similar capacities. In case 2, there has an incremental capacity gap of teachers among four groups of teacher ensembles. $\sigma$ means the standard deviation.

## Comparison of WKD, TAKD and SKD

As mentioned above, the proposed SKD is similar to two multi-teacher KD approaches, i.e., WKD and TAKD, and simultaneously solves their drawbacks (e.g., losing diversity of WKD, sensitiveness for weak-capacity teacher of TAKD). We employ WKD, TAKD and SKD to optimize SKDBERT$_6$ on GLUE-dev and show the GLUE score in Table 5. Moreover, WKD employs five types of sampling probability distribution, i.e., MT-BERT (Wu, Wu, and Huang 2021), uniform[5], teacher-rank and student-rank.

**Diversity Preservation** WKD uses the weighted logits of all teachers to make more comprehensive decision for student distillation, but loses diversity of each teacher (Tran et al. 2020). Besides, TAKD (Mirzadeh et al. 2020) suffers from knowledge vanishing (similar to gradient vanishing (He et al. 2016)) with respect to strong teacher, where the student can only accept few knowledge from the strongest

[4]The improvement is about 0.2% GLUE score for SKDBERT$_6$ with teacher ensembles of T$_{04}$ to T$_{12}$ and T$_{04}$ to T$_{14}$.

[5]This case is identical to AvgKD (Hinton, Vinyals, and Dean 2015).

| Model | GLUE Score (%) | | | | |
|---|---|---|---|---|---|
| | $T_{13}$-$T_{14}$ | $T_{10}$-$T_{14}$ | $T_{07}$-$T_{14}$ | $T_{04}$-$T_{14}$ | $T_{01}$-$T_{14}$ |
| WKD (MT-BERT)† | **86.9** | 86.8 | 86.6 | 86.9 | 86.7 |
| WKD (Uniform) | **86.9** | **87.1** | **86.8** | 86.7 | 86.7 |
| WKD (Teacher-rank) | 86.8 | **87.1** | 86.7 | **87.2** | **86.8** |
| WKD (Student-rank) | 86.7 | 86.8 | 86.6 | 86.6 | **86.8** |
| TAKD | 86.8 | 86.6 | 85.9 | 84.4 | 81.4 |
| SKD (Ours) | **87.1** | **87.5** | **87.4** | **87.6** | **87.1** |

Table 5: Scores of WKD with various sampling distributions, TAKD, SKD for $\text{SKDBERT}_6$ with five groups of teacher ensembles on GLUE-dev. † indicates that the implementation details can be found in Section A of supplementary materials.

teacher. With the sampled one-to-one distillation in each iteration, the student can preserve the diversity of each teacher via SKD to achieve high performance. In five groups of teacher ensemble, SKD outperforms both WKD and TAKD. In particular, SKD achieves 0.6% higher GLUE score than the best one of WKD and TAKD using the teacher ensemble of $T_{07}$ to $T_{14}$.

**Sensibility to Teachers' Capacity** For WKD, weak-capacity teacher hardly reduce its performance as shown in Table 5. However, TAKD is very sensitive to the capacity of weakest teacher. With the capacity reduction of the weakest teacher in the teacher ensemble, the performance of TAKD deteriorates dramatically. For instance, TAKD with $T_{01}$ to $T_{14}$ is 5.4% worse than TAKD with $T_{13}$ to $T_{14}$. The proposed SKD can avoid limiting the performance of the student to the weakest teacher. For example, SKDBERT utilizes the teacher ensemble of $T_{01}$ to $T_{14}$ to achieve 87.1% GLUE score which outperforms all teachers except the strongest $T_{14}$.

Above all, SKD can not only preserve the diversity of each teacher, but also fill the large capacity gap between the teacher and the student while avoiding limiting the performance of the student to the weakest teachers.

## Impact of Extra Learning Procedure and Distillation Objective

To obtain high-performance BERT-style language model, TinyBERT (Jiao et al. 2020) employs extra learning procedure (i.e., Data Augmentation (DA) with GloVe word embedding (Pennington, Socher, and Manning 2014)) and distillation objective (i.e., Transformer Distillation (TD) including transformer-layer distillation and embedding-layer distillation). For a fair comparison with TinyBERT, we verify the effectiveness of the combination of SKD with DA and TD, and show the results on GLUE-test in Table 6.

As shown in Table 6, SKDBERT with the technique of DA achieves the best performance on three out of four tasks compared to TinyBERT. However, TD shows unsatisfactory performance for both SKDBERT and TinyBERT. For SKDBERT, we consider the possible reason of this phenomenon is that we employ vanilla KD with $\text{BERT}_{\text{BASE}}$ rather than SKD with a specific teacher ensemble for TD, due to the different number of layer and hidden size of teachers. As a result, this leads to a knowledge gap between transformer distillation with $\text{BERT}_{\text{BASE}}$ and prediction layer distillation with SKD. Unfortunately, the above mentioned reason is not appropriate for TinyBERT whose performance can be improved by both TD and DA (Jiao et al. 2020). We consider the possible reason of this phenomenon is that the employed teacher is different.

| Model | Teacher | DA | TD | MRPC | RTE | SST-2 | QNLI | Avg |
|---|---|---|---|---|---|---|---|---|
| $\text{TinyBERT}_6$† | $T_{12}$ | ✗ | ✗ | 87.6 | 65.6 | 91.4 | 90.4 | 83.8 |
| | | ✓ | ✗ | **88.6** | 66.8 | 92.1 | 90.6 | 84.5 |
| | | ✗ | ✓ | 86.7 | 61.8 | 92.0 | 89.7 | 82.6 |
| | | ✓ | ✓ | 87.9 | 62.9 | 92.9 | 90.2 | 83.5 |
| $\text{SKDBERT}_6$ | $T_{07}$-$T_{12}$ | ✗ | ✗ | 88.5 | 67.2 | 91.9 | 89.9 | 84.4 |
| | | ✓ | ✗ | 88.4 | **67.6** | 93.0 | 90.9 | **85.0** |
| | | ✗ | ✓ | 87.4 | 62.7 | 93.0 | 90.2 | 83.3 |
| | | ✓ | ✓ | 88.3 | 62.6 | **93.1** | **91.0** | 83.8 |

Table 6: Results of TinyBERT and SKDBERT with extra learning procedure and distillation objective on GLUE-test. DA denotes data augmentation, and TD means transformer distillation. † indicates that these results are obtained by TinyBERT with SKDBERT employed fine-tuned teacher using the code publicly at https://github.com/huawei-noah/Pretrained-Language-Model/tree/master/TinyBERT. MRPC task is evaluated by F1 score, and other tasks are evaluated by accuracy score.

Above all, the combination of SKDBERT and DA can achieve better performance than TinyBERT.

## Conclusion and Future Work

In this paper, we propose a novel distillation paradigm named Stochastic Knowledge Distillation (SKD) to obtain compact BERT-style language model dubbed SKDBERT. SKD samples a teacher from a teacher ensemble, to preserve the diversity of each teacher via the one-to-one distillation, and fill the large capacity gap between the teacher and student while avoid limiting the performance of the student to the weak teachers. Extensive experiments on the GLUE benchmark show that SKDBERT achieves competitive performance while reducing almost half the size and being 100% faster.

However, the efficiency and flexibility of SKDBERT are not enough due to the fixed sampling probability distribution and teacher ensemble. In future work, we will study how to automatically determine the sampling distribution to improve performance and reduce time consumption.

## Acknowledgements

The authors would like to thank Nannan Li, Guangzheng Hu, Jiajun Chai, Yao Shi, Weifan Li, Junjie Wang and Minsong Liu for their thoughtful comments and suggestions, Bear Shi and CC Ding for encouragement and companionship.

## References


Allingham, J. U.; Wenzel, F.; Mariet, Z. E.; Mustafa, B.; Puigcerver, J.; Houlsby, N.; Jerfel, G.; Fortuin, V.; Lakshminarayanan, B.; Snoek, J.; et al. 2021. Sparse MoEs meet efficient ensembles. *arXiv preprint arXiv:2110.03360*.

Bentivogli, L.; Clark, P.; Dagan, I.; and Giampiccolo, D. 2009. The Fifth PASCAL Recognizing Textual Entailment Challenge. In *TAC*.

Cer, D.; Diab, M.; Agirre, E.; Lopez-Gazpio, I.; and Specia, L. 2017. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. In *Proceedings of the 11th Inter- national Workshop on Semantic Evaluation*.

Chen, D.; Li, Y.; Qiu, M.; Wang, Z.; Li, B.; Ding, B.; Deng, H.; Huang, J.; Lin, W.; and Zhou, J. 2020. AdaBERT: Task-adaptive bert compression with differentiable neural architecture search. *arXiv preprint arXiv:2001.04246*.

Chen, Z.; Zhang, H.; Zhang, X.; and Zhao, L. 2018. Quora question pairs. *University of Waterloo*, 1–7.

Clark, K.; Luong, M.-T.; Le, Q. V.; and Manning, C. D. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. 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*, volume 1, 4171–4186.

Dolan, B.; and Brockett, C. 2005. Automatically constructing a corpus of sentential paraphrases. In *Third International Workshop on Paraphrasing, IWP*.

Fan, A.; Grave, E.; and Joulin, A. 2020. Reducing Transformer Depth on Demand with Structured Dropout. In *8th International Conference on Learning Representations, ICLR*.

Guo, F.; Liu, S.; Mungall, F. S.; Lin, X.; and Wang, Y. 2019. Reweighted proximal pruning for large-scale language representation. *arXiv preprint arXiv:1909.12486*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, CVPR*, 770–778.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Iandola, F. N.; Shaw, A. E.; Krishna, R.; and Keutzer, K. W. 2020. SqueezeBERT: What can computer vision teach NLP about efficient neural networks? *arXiv preprint arXiv:2006.11316*.

Jiao, X.; Yin, Y.; Shang, L.; Jiang, X.; Chen, X.; Li, L.; Wang, F.; and Liu, Q. 2020. TinyBERT: Distilling BERT for Natural Language Understanding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP*, 4163–4174.

Kovaleva, O.; Romanov, A.; Rogers, A.; and Rumshisky, A. 2019. Revealing the Dark Secrets of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP*, 4364–4373.

Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; and Soricut, R. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *8th International Conference on Learning Representations, ICLR*.

Li, T.; Li, J.; Liu, Z.; and Zhang, C. 2020. Few sample knowledge distillation for efficient network compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 14639–14647.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Micaelli, P.; and Storkey, A. J. 2019. Zero-shot knowledge transfer via adversarial belief matching. In *Advances in Neural Information Processing Systems, NeurIPS*, volume 32.

Michel, P.; Levy, O.; and Neubig, G. 2019. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems, NeurIPS*, volume 32.

Mirzadeh, S. I.; Farajtabar, M.; Li, A.; Levine, N.; Matsukawa, A.; and Ghasemzadeh, H. 2020. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI Conference on Artificial Intelligence, AAAI*, volume 34, 5191–5198.

Pan, H.; Wang, C.; Qiu, M.; Zhang, Y.; Li, Y.; and Huang, J. 2021. Meta-KD: A Meta Knowledge Distillation Framework for Language Model Compression across Domains. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP*, 3026–3036.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP*, 1532–1543.

Polino, A.; Pascanu, R.; and Alistarh, D. 2018. Model compression via distillation and quantization. In *6th International Conference on Learning Representations, ICLR*.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21: 140:1–140:67.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100, 000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP*, 2383–2392.

Sajjad, H.; Dalvi, F.; Durrani, N.; and Nakov, P. 2020. Poor man's bert: Smaller and faster transformer models. *arXiv preprint arXiv:2004.03844*.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Shen, S.; Dong, Z.; Ye, J.; Ma, L.; Yao, Z.; Gholami, A.; Mahoney, M. W.; and Keutzer, K. 2020. Q-BERT: Hessian based ultra low precision quantization of BERT. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 8815–8821.

Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C. D.; Ng, A. Y.; and Potts, C. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP*, 1631–1642.

Sun, S.; Cheng, Y.; Gan, Z.; and Liu, J. 2019. Patient Knowledge Distillation for BERT Model Compression. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP*, 4322–4331.

Sun, Z.; Yu, H.; Song, X.; Liu, R.; Yang, Y.; and Zhou, D. 2020. MobileBERT: A Compact Task-Agnostic BERT for Resource-Limited Devices. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL*, 2158–2170.

Tran, L.; Veeling, B. S.; Roth, K.; Swiatkowski, J.; Dillon, J. V.; Snoek, J.; Mandt, S.; Salimans, T.; Nowozin, S.; and Jenatton, R. 2020. Hydra: Preserving ensemble diversity for model distillation. In *International Conference on Machine Learning Workshop on Uncertainty and Robustness in Deep Learning*.

Turc, I.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Well-read students learn better: The impact of student initialization on knowledge distillation. *arXiv preprint arXiv:1908.08962*.

Voita, E.; Talbot, D.; Moiseev, F.; Sennrich, R.; and Titov, I. 2019. Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL*, 5797–5808.

Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. R. 2019. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *7th International Conference on Learning Representations, ICLR*.

Wang, W.; Wei, F.; Dong, L.; Bao, H.; Yang, N.; and Zhou, M. 2020. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems, NeurIPS*, volume 33, 5776–5788.

Williams, A.; Nangia, N.; and Bowman, S. R. 2017. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2017 Conference of the North American Chapter of the Association for Computational Linguistics, NAACL*.

Wu, C.; Wu, F.; and Huang, Y. 2021. One teacher is enough? pre-trained language model distillation from multiple teachers. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP*, 4408–4413.

Xie, Q.; Luong, M.-T.; Hovy, E.; and Le, Q. V. 2020. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 10687–10698.

Xu, C.; Zhou, W.; Ge, T.; Wei, F.; and Zhou, M. 2020. BERT-of-Theseus: Compressing BERT by Progressive Module Replacing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP*, 7859–7869.

Xu, J.; Tan, X.; Luo, R.; Song, K.; Li, J.; Qin, T.; and Liu, T.-Y. 2021. NAS-BERT: Task-agnostic and adaptive-size BERT compression with neural architecture search. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 1933–1943.

Xu, J.; Tan, X.; Song, K.; Luo, R.; Leng, Y.; Qin, T.; Liu, T.-Y.; and Li, J. 2022. Analyzing and mitigating interference in neural architecture search. In *International Conference on Machine Learning, ICML*, 24646–24662. PMLR.

Yang, C.; An, Z.; Cai, L.; and Xu, Y. 2021. Hierarchical self-supervised augmented knowledge distillation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI*, 1217–1223.

Yang, C.; Zhou, H.; An, Z.; Jiang, X.; Xu, Y.; and Zhang, Q. 2022. Cross-image relational knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12319–12328.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R. R.; and Le, Q. V. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems, NeurIPS*, volume 32.

Yao, H.; Zhang, C.; Wei, Y.; Jiang, M.; Wang, S.; Huang, J.; Chawla, N.; and Li, Z. 2020. Graph few-shot learning via knowledge transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence, AAAI*, volume 34, 6656–6663.

Yuan, F.; Shou, L.; Pei, J.; Lin, W.; Gong, M.; Fu, Y.; and Jiang, D. 2021. Reinforced multi-teacher selection for knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence, AAAI*.

Zagoruyko, S.; and Komodakis, N. 2016. Wide residual networks. In *Proceedings of the British Machine Vision Conference 2016, BMVC*.