Minimal Distillation Schedule for Extreme Language Model Compression

Chen Zhang[♣], **Yang Yang**[♦], **Qifan Wang**[♥], **Jiahao Liu**[♦], **Jingang Wang**[♦], **Wei Wu**[♦], **Dawei Song**[♣♠*]

[♣]Beijing Institute of Technology [♦]Meituan NLP [♥]Meta AI [♠]The Open University

chenzhang9702@outlook.com

Abstract

Recent studies have revealed that language model distillation can become less effective when there is a significant capacity gap between the teacher and the student models. In order to bridge the gap, teacher assistant-based distillation has been introduced, in which the selection of the teacher assistant plays a crucial role in transferring knowledge from the teacher to the student. However, existing approaches for teacher assistant-based distillation require numerous trials to find the optimal teacher assistant. In this paper, we propose a novel approach called Minimal Distillation Schedule (MINIDISC), which enables the scheduling of an optimal teacher assistant in just one trial for extreme model compression (e.g, to 5% scale). In particular, we empirically show that the performance of the student is positively correlated with the scale-performance tradeoff of the teacher assistant. We then introduce a new λ-tradeoff metric that quantifies the optimality of the teacher assistant without the need for trial distillation to the student. By employing a sandwich framework, MINIDISC can select the optimal teacher assistant with the best λ-tradeoff. We extensively evaluate MINIDISC through a series of experiments on the GLUE benchmark. The results demonstrate that our approach achieved an improved efficiency compared to various state-of-the-art baselines. Furthermore, we showcase the scalability of MINIDISC by applying it to a language model with billions of parameters.[1]

1 Introduction

Pretrained language models (LMs) (Devlin et al., 2019; Liu et al., 2019; Radford et al., 2019; Brown et al., 2020; Raffel et al., 2020) have achieved promising results in various downstream tasks (Wang et al., 2019; Rajpurkar et al., 2018),

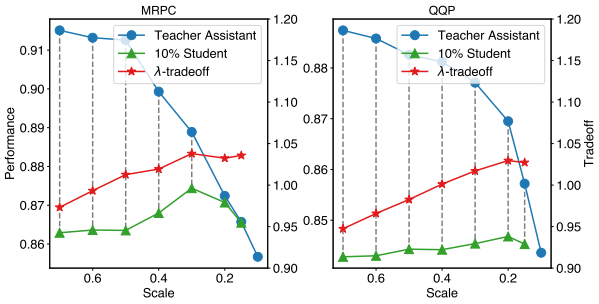


Figure 1: The impact of teacher assistants of different scales and performance on the performance of students. In the study, a $\text{BERT}_{\text{base}}$ model is used as the teacher and distilled to a pruned student (10% parameters of the teacher) via different teacher assistants (Mirzadeh et al., 2020) on MRPC and QQP. There are several observations: (1) The blue curve shows that the performance of the teacher assistant degrades with the decreasing of its scale, which is obvious. (2) The green curve validates that the performance of the student varies with different teacher assistants. (3) The red curve represents λ-tradeoff of the teacher assistant, which is positively correlated with the performance of the student.

but are inapplicable to those requiring limited computational resources (Liu et al., 2021b). To address this issue, LMs can be compressed using a range of strategies such as model quantization (Zafrir et al., 2019; Bai et al., 2021), pruning (Michel et al., 2019; Hou et al., 2020), etc., among which knowledge distillation (Sun et al., 2019; Wang et al., 2020) has gained significant attention. It operates within the teacher-student framework, where a large model acts as the teacher, transferring its knowledge to a smaller student model.

Recent advances (Mirzadeh et al., 2020) have shown a significant performance decline in conventional distillation methods when dealing with a substantial capacity gap between the teacher and the student models. To alleviate this, teacher assistant-based distillation (Son et al., 2021) has been proposed. This approach involves distilling the teacher model into an intermediate-scale teacher assistant,

*Corresponding author.

[1]The code is available at https://github.com/GeneZC/MiniDisc.

which then serves as an intermediary to transfer knowledge to the student model. While teacher assistant-based distillation generally lifts the performance of the student (Wang et al., 2020; Wu et al., 2021), the performance of the student is largely impacted by the choice of the teacher assistant as illustrated in Figure 1. In fact, we observe there is potentially a turning point of the student performance, indicating a scale-performance (i.e., x- v.s. y-axis) tradeoff in scheduling the teacher assistant. However, existing studies schedule the teacher assistant in an enumeration manner, resulting in an inferior solution that requires maximally many trials to meet the optimal teacher assistant (maximal distillation schedule, in short MAXIDISC).

To this demand, we propose a minimal distillation schedule (MINIDISC) that enables the identification of the optimal teacher assistant in just a single trial. We define a *λ-tradeoff* metric to empirically measure the tradeoff between scale and performance for a given teacher assistant, as depicted in Figure 1. This allows us to determine the optimality of the teacher assistant without requiring multiple trial distillations to the student model. To efficiently obtain the optimal teacher assistant based on the *λ-tradeoff* metric, we introduce MINIDISC within a sandwich framework, consisting of three stages. In the *specification* stage, we utilize gridding and pruning techniques to generate a series of teacher assistant candidates with varying scales. In the *optimization* stage, we demonstrate that the generated candidates adhere to the incremental property and the sandwich rule. Furthermore, we present two approximations that enable the computation of the *λ-tradeoff* for each teacher assistant candidate at a lower computational cost. In the *selection* stage, we choose the optimal teacher assistant by selecting the candidate with the highest λ-tradeoff value. It is worth noting that MINIDISC can be directly extended to scenarios involving multiple sequential teacher assistants by recursively applying the MINIDISC procedure. However, this work focuses on a single teacher assistant as it is sufficiently effective.

To verify the effectiveness of MINIDISC, we conduct experiments on GLUE (Wang et al., 2019). Experimental results exhibit the competitive performance of MINIDISC compared to several state-of-the-art baselines, with improved efficiency (10×) of MINIDISC compared to MAXIDISC. Further, MINIDISC is applied to large LMs EncT5_{xl} (Liu et al., 2021a) and $\text{LLaMA2}_{\text{7B}}$ (Touvron et al., 2023) to show its scalability.

2 Related Work

Model Pruning Model pruning (Han et al., 2015) spans from unstructured pruning (Frankle and Carbin, 2019; Louizos et al., 2018; Sanh et al., 2020; Chen et al., 2020) to structured pruning (Michel et al., 2019; Hou et al., 2020; Li et al., 2017; Xia et al., 2022; Lagunas et al., 2021). Unstructured pruning prunes parameters at neuron level referring to parameter magnitude (Han et al., 2015; Louizos et al., 2018) or learning dynamics (Sanh et al., 2020), while structured pruning (Michel et al., 2019; Xia et al., 2022) prunes parameters at module level relying on parameter sensitivity. Although unstructured pruning enjoys a finer-grained pruning, it can only fit specialized devices. In contrast, structured pruning generally fits modern acceleration devices. In our work, we adopt structured pruning for deriving the structures of candidates for its benefits for distillation. Pruning also offers an opportunity to optimize the efficiency and effectiveness of our method due to its merits (Li et al., 2017; Frankle and Carbin, 2019; Yu and Huang, 2019; Cai et al., 2020; Liang et al., 2021; Ma et al., 2022; Yang et al., 2022b,a).

Knowledge Distillation Knowledge distillation (Hinton et al., 2015) can be divided into two categories: task-specific (Sun et al., 2019; Hinton et al., 2015; Li et al., 2020; Park et al., 2021) and task-agnostic (Wang et al., 2020; Turc et al., 2019; Sanh et al., 2019; Sun et al., 2020; Jiao et al., 2020; Wang et al., 2021) distillation. Task-specific methods distill finetuned models with task-specific data, while task-agnostic methods distill pretrained models directly with task-agnostic data. Learning bjective is central to distillation, and distilling logits (Hinton et al., 2015) is the most common way. Recently, hidden states (Sanh et al., 2019; Sun et al., 2020), attention distributions (Jiao et al., 2020; Wang et al., 2020; Li et al., 2020; Wang et al., 2021), and high-order relations (Park et al., 2021) are taken into consideration for better abstraction. Teacher assistant-based distillation (Wang et al., 2020; Mirzadeh et al., 2020; Wu et al., 2021) is showcased to trade in teacher scale for student performance by inserting an intermediate teacher assistant. However, setting an optimal teacher assistant for the student is nontrivial. In this work, we aim to achieve this goal.

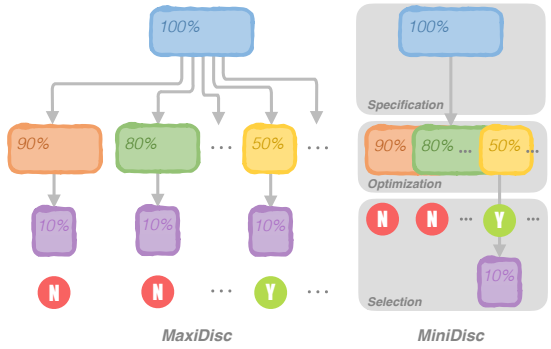


Figure 2: An overview of MINIDISC by contrasting it to MAXIDISC, where one arrow denotes a distillation step. MINIDISC uses only one trial while MAXIDISC uses many trials to schedule the optimal teacher assistant.

3 Methodology

3.1 Problem Definition

Given a teacher model $\mathcal{T}$, our goal is to identify an optimal teacher assistant $\mathcal{A}$, such that the performance of the student $\mathcal{S}$ can be maximized when distilling the teacher to the student via the teacher assistant (i.e., $\mathcal{T} \rightarrow \mathcal{A} \rightarrow \mathcal{S}$). Formally, the teacher model is denoted as $(\mathcal{T}, s_t, m_t)$, where s_t and m_t are the **s**acle and perfor**m**ance of the teacher respectively. Similarly, the teacher assistant and the student are denoted as $(\mathcal{A}, s_a, m_a)$ and $(\mathcal{S}, s_s, m_s)$ It is straightforward that the scale and the performance of the teacher assistant are bounded by the teacher and the student.

The overview of MINIDISC is presented in Figure 2. Our MINIDISC uses only one trial while MAXIDISC uses many trials to schedule the optimal teacher assistant. There are three key components in MINIDISC. *Specification*: the scales and structures of candidates are specified by gridding the scale and pruning the structure of the teacher. *Optimization*: candidates are sub-sampled and assembled into a sandwich-like model, thus jointly optimized in the *sandwich framework*. *Selection*: the candidate with the best λ-*tradeoff* is selected, thus the student is distilled in one trail.

3.2 Scale-performance Tradeoff

While the scale-performance tradeoff can be an indicator of a good teacher assistant, it is not easy to measure. To empirically quantify the scale-performance balance, we introduce a new tradeoff measure below:

Definition 1 (***λ-tradeoff***) *The* λ*-tradeoff measure of a teacher assistant* $(\mathcal{A}, s_a, m_a)$ *is defined as* $t_a = m_a + \lambda \cdot (1 - s_a)$*, where* $\lambda \in [0, 1]$.

In practice, we observe that the λ-*tradeoff* (red curves) of the teacher assistant is positively correlated with the performance of the student (green curves). Theoretically, due to the linear property of the λ-*tradeoff* and the concave property of the teacher assistant scale-performance correlation, there should always be one and only one maximum value of λ-*tradeoff*.

3.3 Sandwich Framework

The problem can be reformulated as finding an optimal teacher assistant that has the maximum value of λ-*tradeoff*:

$$
\begin{aligned}
(\mathcal{A}^*, s_a^*, m_a^*) &= \operatorname*{argmax}_{\mathcal{A}, s_a, m_a} t_a \\
&= \underbrace{\underbrace{\operatorname*{argmax}_{s_a} \operatorname*{argmax}_{\mathcal{A}}}_{\text{specification}} \underbrace{\operatorname*{argmax}_{m_a}}_{\text{optimization}} t_a}_{\text{selection}}
\end{aligned} \tag{1}
$$

Based on the above reformulation, a sandwich framework can be implemented to solve the problem with three main stages: *specification*, *optimization*, and *selection*. Essentially, during *specification*, a set of teacher assistant candidates are generated of different scales. Then the performance metric of the teacher assistant of each scale is obtained through an efficient *optimization*. These two stages form a feasible region for the above reformulation. Finally, the optimal teacher assistant $\mathcal{A}^*$ is selected with a linear scanning of the feasible region during *selection*. After the discovery of the optimal teacher assistant, the teacher assistant can subsequently be distilled to the expected student.

Specification We use gridding and pruning techniques to identify the structure of each candidate.

Gridding. Theoretically, one needs to generate candidates at every possible scale to find the optimal solution. However, it is impossible to enumerate all possibilities in a continuous space. Therefore, we discretize the candidate scales into n discrete values, $\{\mathcal{A} = (\mathcal{A}_k, s_{a_k}, m_{a_k}) \mid \Delta s_a = (s_t - s_s)/n\}$, with equal slicing between the teacher scale and student scale.

Pruning. For candidates at various scales, there are still an infinite number of possible structures, e.g., different combinations of width and depth. A number of approaches have been proposed to identify a good structure at a scale, including dynamic search (Hou et al., 2020), layer dropping (Fan et al., 2020) and pruning (Michel et al., 2019). In this work, we adopt pruning to assign structures $\mathcal{A}_k$

to the candidates due to its known advantages in knowledge distillation (Xia et al., 2022). Concretely, following previous work (Michel et al., 2019), the pruning starts with the least important parameters based on their importance scores, which are approximated by masking the parameterized structures. The technical details of our pruning are supplied in Appendix A.

Essentially, gridding positions the scales of candidates between the scales of the teacher and student with equal intervals and pruning assigns candidates with pruned structures.

Optimization A straightforward solution to unearth the optimality of each candidate is exhaustively measuring the student performance distilled from each, e.g., MAXIDISC. *λ-tradeoff* offers a chance to measure the optimality without actual distillation. However, the memory footprints and computational costs apparently can also be extremely large considering the number of candidates when obtaining performance (i.e., m_a) of all candidates. To reduce the memory overhead and the computational complexity, we introduce two effective approximations, *parameter-sharing* and *sandwich-optimization*, so that the *λ-tradeoff*s of all candidates at different scales can be yielded in one run. The feasibility of the approximations are guarded by the following two properties.

Property 1 (Incremental Property) *For two candidates $\mathcal{A}_i$ and $\mathcal{A}_j$ in the teacher assistant candidate set $\mathcal{A}$, if $s_i < s_j$, then we have $\mathcal{A}_i \subset \mathcal{A}_j$.*

This incremental property is an outcome of the pruning approach (Li et al., 2017; Frankle and Carbin, 2019), which essentially tells that among all candidates obtained from the specification, the structure of a candidate at a smaller scale is a subset of the structure for a candidate at a larger scale.

Remark 1 *The incremental property affirms that a larger candidate can result in a smaller one by continuously pruning less significant parameters, which enables these candidates to be assembled into one sandwich-like model in a parameter-sharing fashion. The memory scale of the sandwich-like model is exactly that of the largest candidate.*

Property 2 (Sandwich Rule) *For two candidates $\mathcal{A}_i$ and $\mathcal{A}_j$ from candidate set $\mathcal{A}$, if $s_i < s_j$, then we have $m_s \leq m_i \leq m_j \leq m_t$.*

The sandwich rule (Yu and Huang, 2019; Cai et al., 2020) states that the performance of a candidate is bounded by the best performance of a larger candidate and a smaller one, due to the subset structure. Therefore, a candidate can be optimized by alternatively distilling its larger and smaller candidates, without direct distillation.

Remark 2 *The sandwich rule allows us to subsample η out of all n ($\eta \leq n$) filling-like candidates and conduct sandwich-optimization over the sampled candidates, which substantially reduces the computational cost.*

With the two approximations, we reduce the memory footprints of all candidates to a distinguished one via parameter-sharing. The computational costs are also largely reduced with sandwich-optimization. Finally, we formulate the distillation objectives for task-specific distillation (TSD) and task-agnostic distillation (TAD) respectively as:

$$\begin{aligned}\mathcal{L}_{\text{TSD}} &= \sum_{i=1}^{\eta} \text{CE}(\mathbf{y}_{\mathcal{T}}, \mathbf{y}_{\mathcal{A}_i}) + \text{MSE}(\mathbf{H}_{\mathcal{T}}, \mathbf{H}_{\mathcal{A}_i}) \\ \mathcal{L}_{\text{TAD}} &= \sum_{i=1}^{\eta} \text{KL}(\mathbf{R}^{\mathsf{Q}}_{\mathcal{T}}, \mathbf{R}^{\mathsf{Q}}_{\mathcal{A}_i}) + \text{KL}(\mathbf{R}^{\mathsf{K}}_{\mathcal{T}}, \mathbf{R}^{\mathsf{K}}_{\mathcal{A}_i}) \\ &\quad + \text{KL}(\mathbf{R}^{\mathsf{V}}_{\mathcal{T}}, \mathbf{R}^{\mathsf{V}}_{\mathcal{A}_i})\end{aligned} \tag{2}$$

where MSE, CE and KL stand for mean squared error, cross entropy and kullback-leibler divergence respectively. $\mathbf{H}$ is the last layer of hidden states, $\mathbf{y}$ is the final prediction. As is taken from MiniLM (Wang et al., 2021), $\mathbf{R}^{\mathsf{Q}}$ is the query relation matrix containing totally h attention heads from the last layer, likewise $\mathbf{R}^{\mathsf{K}}$ and $\mathbf{R}^{\mathsf{V}}$ are the key and value relation matrices. Since heads can be pruned for a teacher assistant candidate, an additional self-attention module is employed as the last layer for TAD. The teacher assistants with the best performance at different scales can be obtained after the above optimization. The unsampled teacher assistants can be retrieved based on the larger teacher assistant from the sampled pool using the shared parameters.

Selection The optimal teacher assistant can be identified by selecting the candidate with the best *λ-tradeoff* measure, which is then distilled to the expected student again following above distillation objectives. Note that the tradeoff measure is also dependent on λ. However, we empirically find that the optimal solution of MINIDISC is relatively stable with a wide range of λ, and we fix λ to 0.2 in all our experiments. More discussion on the impact of λ is provided in the experiments.

4 Experiments

4.1 Setup

Datasets and Metrics We conduct experiments on GLUE (Wang et al., 2019). The GLUE originally consists of two sequence classification tasks, SST-2 (Socher et al., 2013) and CoLA (Warstadt et al., 2019), with seven sequence-pair classification tasks, i.e., MRPC (Dolan and Brockett, 2005), STS-B (Cer et al., 2017), QQP, MNLI (Williams et al., 2018), QNLI (Rajpurkar et al., 2016), RTE (Bentivogli et al., 2009) and WNLI (Levesque et al., 2012). We exclude WNLI and CoLA due to the evaluation inconsistency (in other words, compressed LMs get dramatically worse results while original LMs get much better ones as found out in (Xia et al., 2022)) and use the other seven tasks for evaluation. Following the work in BERT (Devlin et al., 2019), we report F1 on MRPC and QQP, Spearman Correlation scores (Sp Corr) on STS-B, and Accuracy (Acc) on other tasks. Macro average scores (Average) over these seven tasks are computed for overall performance. Results on development sets are reported. We also adopt Wikipedia for pretraining in task-agnostic distillation. The detailed statistics, maximum sequence lengths, and metrics of GLUE and Wikipeida are supplied in Appendix B.

Implementation Details Experiments are carried out on $\text{BERT}_{\text{base}}$ (Devlin et al., 2019) and EncT5_{xl} (Liu et al., 2021a). EncT5 is a language model which achieves competitive performance as T5 (Raffel et al., 2020) on GLUE with a nearly encoder-only T5 (incorporated with a decoder layer). Our task-specific experiments are carried out on either one Nvidia A100 for EncT5_{xl} or one Nvidia V100 for $\text{BERT}_{\text{base}}$, and η is set to 6 according to our empirical investigation. On the other hand, the task-agnostic experiments are carried out on eight Nvidia A100s with $\text{BERT}_{\text{base}}$. η is set to 3 to substantially reduce computational burden. The number of relation heads is set to 32 since we use deep relation distillation as the task-agnostic distillation objective. Other implementation details are supplied in Appendix C. Generally, the sampling is performed from candidates at scales {100%, 95%, 90%, . . . , 10%, 5%}.

Baselines We compare our model with several state-of-the-art baselines. $*_{\text{L};*_{\text{H}}}$ denotes dropping layers and hidden dimensions, while $*_{\%}$ represents structured pruning with either local ranking or our global ranking.

- **Conventional Distillation:** FT (Li et al., 2017) indicates direct finetuning after pruning. KD (Hinton et al., 2015), PKD (Sun et al., 2019) and CKD (Park et al., 2021) are methods with different objectives, i.e., KD directly distills logits, PKD distills both logits and hidden states and CKD distills token and layer relations. DynaBERT (Hou et al., 2020) uses structured pruning with a local ranking in each layer. StarK (Yang et al., 2022a) views sparse teachers as student-friendly teachers. MiniLM (Wang et al., 2021) is distilled with the deep relation alignment. TinyBERT (Jiao et al., 2020) is distilled with a combination of various feature distillations.

- **Teacher Assistant-based Distillation:** TA (Mirzadeh et al., 2020; Wang et al., 2020) is specifically incorporated for both task-specific and task-agnostic distillation with a 40%-scale teacher assistant. MAXIDISC goes further upon TA and manually selects the best teacher assistant among available trials.

4.2 Main Results

Results of Task-specific Distillation Table 1 presents the comparison results of different methods on task-specific distillation at three student scales. There are several key observations: **First**, both MINIDISC and MAXIDISC yield better performance than TA does and MINIDISC obtains similar or even better results compared to MAXIDISC with much fewer GPU hours. This validates the efficiency of MINIDISC for identifying a good teacher assistant. Notably, the slight performance improvement is attributed to parameter sharing, which is detailed in later analysis. For further smaller $\text{BERT}_{3\%}$, the result still holds, as supplied in Appendix D. Additional comparisons of practical inference measurement are supplied in Appendix E. **Second**, pruning based models perform much better compared to the layer dropping methods, e.g., $\text{KD}_{15\%}$ achieves much higher score than FLOPs-matched $\text{KD}_{2\text{L}}$, which verifies the effectiveness of pruning approach in knowledge distillation. Moreover, we discover the global ranking strategy surpasses the local ranking one by comparing $\mathcal{L}_{\text{TSD}15\%}$ to FLOPs-matched $\text{DynaBERT}_{15\%}$. We speculate the structures induced by the local

Table 1: The results of task-specific distillation upon $\text{BERT}_{\text{base}}$. The GPU hours of teacher assistant-based methods are estimated with respect to their conventional counterparts.

Method	FLOPs	SST-2	MRPC	STS-B	QQP	MNLI-m/mm	QNLI	RTE	Average	GPUs
$\text{BERT}_{\text{base}}$	10.9G	93.8	91.5	87.1	88.4	84.9/84.9	91.9	71.5	86.7	–
Conventional Distillation										
KD_{2L} (2015)	1.8G	86.8	82.5	46.8	83.7	73.5/73.1	79.6	58.1	73.0	1×
PKD_{2L} (2019)	1.8G	86.7	82.4	46.8	83.7	73.4/73.0	79.7	57.4	72.9	1×
CKD_{2L} (2021)	1.8G	86.4	82.3	48.6	83.6	73.3/73.0	79.1	56.7	72.9	1×
StarK_{2L} (2022a)	1.8G	88.1	83.1	48.6	83.8	73.9/74.3	80.4	57.8	73.7	1×
$\text{DynaBERT}_{\text{15\%}}$ (2020)	2.2G	89.1	85.1	84.7	84.3	78.3/79.0	86.6	61.4	81.1	1×
$\text{FT}_{\text{15\%}}$ (2017)	1.6G	89.9	87.1	85.6	86.1	79.9/80.1	85.7	63.9	82.3	1×
$\text{KD}_{\text{15\%}}$ (2015)	1.6G	89.9	88.6	85.1	86.2	79.8/80.2	85.6	63.9	82.4	1×
$\mathcal{L}_{\text{TSD15\%}}$	1.6G	90.1	88.9	85.1	86.5	80.0/80.2	86.0	65.3	**82.8**	1×
$\text{FT}_{\text{10\%}}$ (2017)	1.1G	88.2	84.8	84.7	84.4	77.6/77.3	84.3	65.3	80.8	1×
$\text{KD}_{\text{10\%}}$ (2015)	1.1G	88.2	87.6	84.0	84.4	77.6/77.4	84.3	67.2	81.3	1×
$\mathcal{L}_{\text{TSD10\%}}$	1.1G	88.8	87.8	84.0	84.6	77.6/77.5	84.9	66.4	**81.5**	1×
$\text{FT}_{\text{5\%}}$ (2017)	0.5G	85.4	82.8	84.1	82.6	72.5/73.3	81.7	63.9	78.3	1×
$\text{KD}_{\text{5\%}}$ (2015)	0.5G	85.6	84.0	83.8	82.5	72.6/73.2	81.6	63.2	78.3	1×
$\mathcal{L}_{\text{TSD5\%}}$	0.5G	85.4	85.5	83.9	82.7	73.0/73.4	82.7	63.2	**78.7**	1×
Teacher Assistant-based Distillation										
$\text{TA}_{\text{15\%}}$ (2020)	1.6G	89.3	87.7	85.3	85.7	80.0/80.3	88.1	68.4	83.1	2×
$\text{MAXIDISC}_{\text{15\%}}$	1.6G	89.8	87.7	85.4	86.9	81.0/80.1	86.1	68.2	83.2	40×
$\text{MINIDISC}_{\text{15\%}}$	1.6G	89.8	88.2	85.8	86.6	80.3/79.9	87.3	68.2	**83.3**	4×
$\text{TA}_{\text{10\%}}$ (2020)	1.1G	89.1	87.9	83.1	84.7	77.8/77.9	85.7	68.6	81.8	2×
$\text{MAXIDISC}_{\text{10\%}}$	1.1G	89.0	88.2	84.8	84.8	78.3/77.8	85.3	66.8	81.9	40×
$\text{MINIDISC}_{\text{10\%}}$	1.1G	89.1	88.4	85.4	84.9	78.2/78.6	86.3	68.2	**82.4**	4×
$\text{TA}_{\text{5\%}}$ (2020)	0.5G	86.5	86.5	82.2	83.2	73.3/73.7	82.6	65.3	79.2	2×
$\text{MAXIDISC}_{\text{5\%}}$	0.5G	86.9	88.3	84.8	83.7	74.4/76.3	83.5	65.0	**80.4**	40×
$\text{MINIDISC}_{\text{5\%}}$	0.5G	86.9	87.6	84.8	83.5	72.7/74.5	84.0	66.8	80.1	4×

ranking strategy are not that effective. The distribution of example pruned structures is supplied in Appendix F. **Third**, conventional distillation methods generate reasonable results at large student scale but fail to maintain the student performance at small scale. Nonetheless, TA consistently outperforms the conventional baselines at all scales.

Results of Large-scale Distillation As is shown in Table 2, we conduct a similar comparison on a large LM, EncT5_{xl}, with over one billion parameters. The very first results of the large LM also exhibit an akin trend as the one in $\text{BERT}_{\text{base}}$. The results on a more recent large LM $\text{LLaMA2}_{\text{7B}}$ are displayed in Table 3. And the results on a moderate $\text{BERT}_{\text{large}}$ are supplied in Appendix G. We therefore conclude that the scalability of MINIDISC is also compelling. Reversely, the results of MINIDISC on small LMs are supplied in Appendix H.

Results of Task-agnostic Distillation We also apply MINIDISC to task-agnostic distillation and report the results in Table 4. The first glimpse is that $\mathcal{L}_{\text{TAD}}$ surpasses $\mathcal{L}_{\text{TSD}}$, indicating the deep relation alignment is more suitable for task-agnostic distillation. Surprisingly, we discover that the pruned structures can boost the performance of MiniLM, i.e., $\mathcal{L}_{\text{TAD}}$, and establish a new state-of-the-art for conventional task-agnostic distillation. Another interesting observation is that teacher assistant-based distillation methods do not improve the performance over conventional distillation methods until the scale is reduced to 5%, indicating that conventional distillation methods are already promising choices on task-agnostic distillation at large scales. Nonetheless, we still argue the applicability of MINIDISC to task-agnostic distillation for a performance guarantee. Note that the results of TinyBERT with additional task-specific distillation are supplied in Appendix I.

4.3 Analyses

Ablation Study We carry out an ablation study can actually be viewed as a process of bridging MAXIDISC to MINIDISC by firstly adding λ-*tradeoff*, then adding *sandwich framework*. We present the results in Table 5. The results show that:

Table 2: The results of task-specific distillation upon EncT5_{xl}. The GPU hours of teacher assistant-based methods are estimated with respect to their conventional counterparts.

Method	FLOPs	SST-2	MRPC	STS-B	QQP	MNLI-m/mm	QNLI	RTE	Average	GPUs
EncT5_{xl}	155.9G	96.9	95.1	92.3	90.0	90.7/90.9	95.0	88.5	92.4	–
Conventional Distillation										
$\text{FT}_{10\%}$ (2017)	15.6G	91.6	87.1	86.7	87.9	81.9/87.0	66.1	91.6	83.8	1×
$\text{KD}_{10\%}$ (2015)	15.6G	92.2	86.8	86.6	87.9	83.6/83.8	88.1	63.5	84.1	1×
$\mathcal{L}_{\text{TSD}10\%}$	15.6G	94.5	90.2	87.4	87.9	84.7/84.1	90.8	67.5	**85.9**	1×
$\text{FT}_{5\%}$ (2017)	7.8G	90.1	84.8	84.7	86.5	78.0/78.2	83.9	62.8	81.1	1×
$\text{KD}_{5\%}$ (2015)	7.8G	89.9	85.1	85.4	86.6	79.4/79.6	84.2	55.6	80.7	1×
$\mathcal{L}_{\text{TSD}5\%}$	7.8G	92.9	88.0	83.4	85.4	79.6/80.0	87.0	58.8	**81.9**	1×
Teacher Assistant-based Distillation										
$\text{TA}_{10\%}$	15.6G	94.5	90.7	87.4	88.0	85.2/84.6	91.1	69.3	86.3	2×
$\text{MAXIDISC}_{10\%}$	15.6G	94.6	90.5	88.0	88.1	86.2/85.1	91.5	70.4	86.8	40×
$\text{MINIDISC}_{10\%}$	15.6G	94.6	91.5	87.8	87.3	85.9/85.0	91.1	72.2	**86.9**	4×
$\text{TA}_{10\%}$	7.8G	92.3	88.4	83.7	86.0	80.2/80.5	87.5	56.3	81.9	2×
$\text{MAXIDISC}_{10\%}$	7.8G	93.0	88.0	83.9	86.5	81.2/81.6	88.1	67.5	83.7	40×
$\text{MINIDISC}_{10\%}$	7.8G	93.8	89.8	85.3	86.7	82.9/82.7	89.2	64.6	**84.4**	4×

Table 3: The results of task-specific distillation upon $\text{LLaMA2}_{\text{7B}}$. The Alpaca dataset (Taori et al., 2023) is utilized as the distillation data.

Method	MMLU
$\text{LLaMA2}_{\text{7B}}$	46.0
$\text{KD}_{15\%}$	25.6
$\text{TA}_{15\%}$	26.1
$\text{MAXIDISC}_{15\%}$	26.8
$\text{MINIDISC}_{15\%}$	26.9

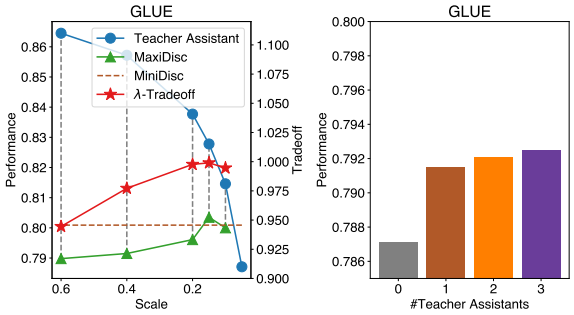


Figure 3: Tradeoff studies by distilling the teacher to a student at 5% scale. On the left hand, the blue curve represents the performance of teacher assistants at different scales. The green curve represents the performance of MAXIDISC using these teacher assistants. The red curve represents the λ-*tradeoff* value. The brown dashed line represents the performance of MINIDISC. On the right hand, the brown, orange, and purple bars represent the performance of MINIDISC using one, two, and three teacher assistants.

1) (MAXIDISC v.s. MAXIDISC w/ λ-*tradeoff*) λ-*tradeoff* can be an accurate measure to select the optimal teacher assistant; 2) (MAXIDISC v.s. MAXIDISC w/ *sandwich framework*) *sandwich framework* can achieve competitive (even slightly better) performance despite the *parameter sharing* among teacher assistant candidates; 3) (MAXIDISC w/ *sandwich framework* v.s. MINIDISC) the two together lead to results slightly better than those of MAXIDISC in a much more efficient manner.

Impact of Candidate Sampling We then study the impact of the *sandwich framework* in MINIDISC by varying the number of sampled candidates η, and measuring the training cost and the student performance. From Table 6, we show the assembled sandwich together with sub-sampled fillings brings acceptable performance detriment and efficiency gain.

Impact of λ To show λ-*tradeoff* is robust on the value of λ, we vary λ within {0.1,0.2,0.3,0.5,0.7}. It can be seen from Table 7 that the performance of MINIDISC is relatively stable with different values of λ. Moreover, we offer a λ-independent solution using a negative derivative of performance to scale as the tradeoff measure, which yields slightly worse results, as supplied in Appendix J.

Existence of Tradeoff To double-check the existence of the concerned tradeoff, we use teacher assistants at different scales within MAXIDISC and plot performance variations of these schedules upon $\text{BERT}_{\text{base}}$ in Figure 3 (left). It can be seen that reducing the teacher assistant scale can

Table 4: The results of task-agnostic distillation upon $\text{BERT}_{\text{base}}$. The results of TinyBERT are reproduced based on their released checkpoints without additional task-specific distillation for a fair comparison. The GPU hours of teacher assistant-based methods are estimated with respect to their conventional counterparts.

Method	FLOPs	SST-2	MRPC	STS-B	QQP	MNLI-m/mm	QNLI	RTE	Average	GPUs
$\text{BERT}_{\text{base}}$	10.9G	93.8	91.5	87.1	88.4	84.9/84.9	91.9	71.5	86.7	–
Conventional Distillation										
$\text{FT}_{10\%}$ (2017)	1.1G	84.6	83.1	83.8	84.5	75.3/75.4	83.2	56.7	78.3	1×
$\mathcal{L}_{\text{TSD}10\%}$	1.1G	90.7	89.0	87.0	85.9	78.4/78.2	86.0	66.4	82.7	1×
$\text{MiniLM}_{4\text{L};384\text{H}}$ (2021)	0.9G	90.0	88.6	87.2	86.1	80.0/80.3	87.9	67.2	83.4	1×
$\mathcal{L}_{\text{TAD}10\%}$	1.1G	92.0	90.1	87.9	86.6	80.0/80.3	88.0	67.2	**84.0**	1×
$\text{FT}_{5\%}$ (2017)	0.5G	84.1	82.4	81.8	83.7	74.4/74.9	82.5	57.0	77.6	1×
$\text{TinyBERT}_{4\text{L};312\text{H}}$ (2020)	0.6G	88.5	87.9	86.6	85.6	78.9/79.2	87.3	67.2	82.7	1×
$\text{MiniLM}_{3\text{L};384\text{H}}$ (2021)	0.7G	89.1	89.1	86.6	85.4	77.8/78.4	87.2	66.1	82.5	1×
$\mathcal{L}_{\text{TAD}5\%}$	0.5G	90.9	89.4	87.7	85.8	79.2/79.8	87.3	65.7	**83.2**	1×
Teacher Assistant-based Distillation										
$\text{TA}_{10\%}$ (2020)	0.9G	90.0	88.5	87.3	86.3	80.1/80.7	88.0	66.4	83.4	2×
$\text{MAXIDISC}_{10\%}$	1.1G	91.5	90.3	87.8	86.6	80.0/80.1	88.6	67.2	**84.0**	40×
$\text{MINIDISC}_{10\%}$	1.1G	91.4	90.0	87.5	86.6	79.8/80.0	88.0	67.2	83.8	4×
$\text{TA}_{5\%}$ (2020)	0.7G	89.8	85.9	86.0	85.5	77.6/78.5	86.8	66.1	82.0	2×
$\text{MAXIDISC}_{5\%}$	0.5G	90.1	89.7	87.4	85.6	79.3/79.7	87.1	67.9	**83.4**	40×
$\text{MINIDISC}_{5\%}$	0.5G	89.3	89.7	87.4	85.9	79.2/79.4	86.9	69.7	**83.4**	4×

Table 5: The ablation study upon distilling $\text{BERT}_{\text{base}}$ to $\text{BERT}_{10\%}$.

Method	GPU hours	MRPC	QQP
$\mathcal{L}_{\text{TSD}10\%}$	1×	87.8	84.6
$\text{MAXIDISC}_{10\%}$	40×	88.2	84.8
w/ λ-*tradeoff*	21×	88.2	84.8
w/ *sandwich framework*	23×	88.4	84.9
$\text{MINIDISC}_{10\%}$	4×	88.4	84.9

Table 6: The impact of candidate sampling upon distilling $\text{BERT}_{\text{base}}$ to $\text{BERT}_{10\%}$.

Method	GPU hours	Average
$\mathcal{L}_{\text{TSD}10\%}$	1×	81.5
$\text{MAXIDISC}_{10\%}$	40×	81.9
$\text{MINIDISC}_{10\%}$ (η=1)	2×	82.1
$\text{MINIDISC}_{10\%}$ (η=3)	2×	81.9
$\text{MINIDISC}_{10\%}$ (η=6)	4×	82.4
$\text{MINIDISC}_{10\%}$ (η=9)	4×	82.4

lead to student performance improvement until a certain scale, after which performance degradation is witnessed. All schedules underperform the λ-*tradeoff* indicated one. We attribute the inferiority to improper scale-performance tradeoffs, as concentrating only on either scale or performance will give rise to a trivial solution with pareto optimality (Sener and Koltun, 2018; Lin et al., 2019). The overall phenomenon implies the existence of scale-performance tradeoff. Similar phenomenon

Table 7: The impact of λ upon distilling $\text{BERT}_{\text{base}}$ to $\text{BERT}_{10\%}$.

Method	MRPC	QQP
$\mathcal{L}_{\text{TSD}10\%}$	87.8	84.6
$\text{MAXIDISC}_{10\%}$	88.2	84.8
$\text{MINIDISC}_{10\%}$ (λ=0.1)	87.5	85.2
$\text{MINIDISC}_{10\%}$ (λ=0.2)	88.4	84.9
$\text{MINIDISC}_{10\%}$ (λ=0.3)	87.5	84.7
$\text{MINIDISC}_{10\%}$ (λ=0.5)	87.8	84.7
$\text{MINIDISC}_{10\%}$ (λ=0.7)	87.8	84.7

is also observed in EncT5, which is supplied in Appendix K.

Sufficiency of One Teacher Assistant To examine whether one teacher assistant is sufficient, we insert more than one teacher assistant to MINIDISC and present the results in Figure 3 (right). It is clear that there is no obvious performance gain when applying more than one teacher assistant (two and three) in schedules. Therefore, we alternatively choose to use only one teacher assistant in MINIDISC for training efficiency based on the sufficiency. The conclusion still holds for EncT5, which is supplied in Appendix K.

Recently proposed progressive distillation methods (Li et al., 2021; Lin et al., 2022), where students are learned firstly from a small teacher then from a larger teacher, inspire us to inspect whether the same regime could further boost MINIDISC

since teacher assistants are essentially small teachers and a natural follow-up action is residually distilling the students from the original teachers (residual distillation). The residual distillation can possibly further improve the performance of MINIDISC, as detailed in Appendix L.

5 Conclusions

In this paper, we propose MINIDISC to identify an optimal teacher assistant for teacher assistant-based distillation in minimally one trial in contrast to MAXIDISC. Having observed that the scale-performance tradeoff of the teacher assistant is of great importance to the performance of the student, we introduce a λ-*tradeoff* measure that quantifies the scale-performance tradeoff of the teacher assistant, and show that it is positively correlated with the student performance. To efficiently compute the measures for teacher assistant candidates and select the optimal one, we design a sandwich optimization for these candidates. Comprehensive results demonstrate the improved efficiency of MINIDISC.

Limitations

Although the value of λ is relatively stable in a wide range, the core limitation of MINIDISC is that the value of λ should be calibrated before practical use. To enable a more automatic process, we conduct some preliminary study by introducing another metric, which does not require any hyper-parameters. More details can be found in Appendix J. We plan to investigate more along this direction in the future. Another limitation of this work is that we leverage gridding and pruning to identify the model structure of each candidate to ensure these candidate structures satisfying certain property for one-run optimization. However, the gridding and pruning process might yield a sub-optimal model architecture at a given model scale. In future, we also plan to explore how to efficient identify an optimal model structure.

Acknowledgements

This work is funded in part by the Natural Science Foundation of China (grant no: 62376027) and Beijing Municipal Natural Science Foundation (grant no: 4222036 and IS23061).

References

Haoli Bai, Wei Zhang, Lu Hou, Lifeng Shang, Jin Jin, Xin Jiang, Qun Liu, Michael R. Lyu, and Irwin King. 2021. Binarybert: Pushing the limit of BERT quantization. In *ACL-IJCNLP*, pages 4334–4348.

Luisa Bentivogli, Bernardo Magnini, Ido Dagan, Hoa Trang Dang, and Danilo Giampiccolo. 2009. The fifth PASCAL recognizing textual entailment challenge. In *TAC*.

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. 2020. Language models are few-shot learners. In *NeurIPS*.

Han Cai, Chuang Gan, Tianzhe Wang, Zhekai Zhang, and Song Han. 2020. Once-for-all: Train one network and specialize it for efficient deployment. In *ICLR*.

Daniel M. Cer, Mona T. Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *SemEval@ACL*, pages 1–14.

Tianlong Chen, Jonathan Frankle, Shiyu Chang, Sijia Liu, Yang Zhang, Zhangyang Wang, and Michael Carbin. 2020. The lottery ticket hypothesis for pre-trained BERT networks. In *NeurIPS*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *IWP@IJCNLP*.

Angela Fan, Edouard Grave, and Armand Joulin. 2020. Reducing transformer depth on demand with structured dropout. In *ICLR*.

Jonathan Frankle and Michael Carbin. 2019. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *ICLR*.

Song Han, Jeff Pool, John Tran, and William Dally. 2015. Learning both weights and connections for efficient neural network. In *NeurIPS*, pages 1135–1143.

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

Lu Hou, Zhiqi Huang, Lifeng Shang, Xin Jiang, Xiao Chen, and Qun Liu. 2020. Dynabert: Dynamic BERT with adaptive width and depth. In *NeurIPS*.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. Tinybert: Distilling BERT for natural language understanding. In *EMNLP*, volume EMNLP 2020 of *Findings of ACL*, pages 4163–4174.

François Lagunas, Ella Charlaix, Victor Sanh, and Alexander M. Rush. 2021. Block pruning for faster transformers. In *EMNLP*, pages 10619–10629.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *KR*.

Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. 2017. Pruning filters for efficient convnets. In *ICLR*.

Jianquan Li, Xiaokang Liu, Honghong Zhao, Ruifeng Xu, Min Yang, and Yaohong Jin. 2020. BERT-EMD: many-to-many layer mapping for BERT compression with earth mover's distance. In *EMNLP*, pages 3009–3018.

Lei Li, Yankai Lin, Shuhuai Ren, Peng Li, Jie Zhou, and Xu Sun. 2021. Dynamic knowledge distillation for pre-trained language models. In *EMNLP*, pages 379–389.

Chen Liang, Simiao Zuo, Minshuo Chen, Haoming Jiang, Xiaodong Liu, Pengcheng He, Tuo Zhao, and Weizhu Chen. 2021. Super tickets in pre-trained language models: From model compression to improving generalization. In *ACL*, pages 6524–6538. Association for Computational Linguistics.

Xi Lin, Hui-Ling Zhen, Zhenhua Li, Qingfu Zhang, and Sam Kwong. 2019. Pareto multi-task learning. In *NeurIPS*, pages 12037–12047.

Zhenghao Lin, Yeyun Gong, Xiao Liu, Hang Zhang, Chen Lin, Anlei Dong, Jian Jiao, Jingwen Lu, Daxin Jiang, Rangan Majumder, and Nan Duan. 2022. PROD: progressive distillation for dense retrieval. *arXiv*, abs/2209.13335.

Frederick Liu, Siamak Shakeri, Hongkun Yu, and Jing Li. 2021a. Enct5: Fine-tuning T5 encoder for non-autoregressive tasks. *CoRR*, abs/2110.08426.

Xiangyang Liu, Tianxiang Sun, Junliang He, Lingling Wu, Xinyu Zhang, Hao Jiang, Zhao Cao, Xuanjing Huang, and Xipeng Qiu. 2021b. Towards efficient NLP: A standard evaluation and A strong baseline. *arXiv*, abs/2110.07038.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *arXiv*, abs/1907.11692.

Christos Louizos, Max Welling, and Diederik P. Kingma. 2018. Learning sparse neural networks through l_0 regularization. In *ICLR*.

Fang Ma, Chen Zhang, Lei Ren, Jingang Wang, Qifan Wang, Wei Wu, Xiaojun Quan, and Dawei Song. 2022. Xprompt: Exploring the extreme of prompt tuning. *CoRR*, abs/2210.04457.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? In *NeurIPS*, pages 14014–14024.

Seyed-Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. 2020. Improved knowledge distillation via teacher assistant. In *AAAI*, pages 5191–5198.

Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, and Jan Kautz. 2017. Pruning convolutional neural networks for resource efficient inference. In *ICLR*.

Geondo Park, Gyeongman Kim, and Eunho Yang. 2021. Distilling linguistic context for language model compression. In *EMNLP*, pages 364–378.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21:140:1–140:67.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *ACL*, pages 784–789.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *EMNLP*, pages 2383–2392.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv*, abs/1910.01108.

Victor Sanh, Thomas Wolf, and Alexander M. Rush. 2020. Movement pruning: Adaptive sparsity by fine-tuning. In *NeurIPS*.

Ozan Sener and Vladlen Koltun. 2018. Multi-task learning as multi-objective optimization. In *NeurIPS*, pages 525–536.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, pages 1631–1642.

Wonchul Son, Jaemin Na, Junyong Choi, and Wonjun Hwang. 2021. Densely guided knowledge distillation using multiple teacher assistants. In *ICCV*, pages 9375–9384.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019. Patient knowledge distillation for BERT model compression. In *EMNLP-IJCNLP*, pages 4322–4331.

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. 2020. Mobilebert: a compact task-agnostic BERT for resource-limited devices. In *ACL*, pages 2158–2170.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: The impact of student initialization on knowledge distillation. *arXiv*, abs/1908.08962.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *ICLR*.

Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. 2021. Minilmv2: Multi-head self-attention relation distillation for compressing pretrained transformers. In *ACL-IJCNLP*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 2140–2151.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *NeurIPS*.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *TACL*, 7:625–641.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL-HLT*, pages 1112–1122.

Yimeng Wu, Mehdi Rezagholizadeh, Abbas Ghaddar, Md. Akmal Haidar, and Ali Ghodsi. 2021. Universal-kd: Attention-based output-grounded intermediate layer knowledge distillation. In *EMNLP*, pages 7649–7661.

Mengzhou Xia, Zexuan Zhong, and Danqi Chen. 2022. Structured pruning learns compact and accurate models. *arXiv*, abs/2204.00408.

Yi Yang, Chen Zhang, and Dawei Song. 2022a. Sparse teachers can be dense with knowledge. *CoRR*, abs/2210.03923.

Yi Yang, Chen Zhang, Benyou Wang, and Dawei Song. 2022b. Doge tickets: Uncovering domain-general language models by playing lottery tickets. In *NLPCC*, volume 13551 of *Lecture Notes in Computer Science*, pages 144–156. Springer.

Jiahui Yu and Thomas S. Huang. 2019. Universally slimmable networks and improved training techniques. In *ICCV*, pages 1803–1811.

Ofir Zafrir, Guy Boudoukh, Peter Izsak, and Moshe Wasserblat. 2019. Q8BERT: quantized 8bit BERT. In *EMC2@NeurIPS*, pages 36–39.

A Technical Details of Pruning

Concretely, following previous work (Michel et al., 2019), the pruning always starts with the least important parameters, which are identified according to importance scores. The importance scores are approximated by first masking the parameterized structures. μ_i, ν_i, and ξ_j denote the mask variables respectively for a self-attention head, optionally a cross-attention head, and a feed-forward neuron, such that for an intermediate input $\mathbf{X}$ and potentially an encoder-produced input $\mathbf{E}$:

$$\begin{aligned}\mathbf{Z} &= \text{SelfAttention}(\mathbf{X}) \\ &= \sum_i^h \mu_i \cdot \text{softmax}(\mathbf{X}\mathbf{W}_i^{\mathsf{Q}}\mathbf{W}_i^{\mathsf{K}\top}\mathbf{X}^\top)\mathbf{X}\mathbf{W}_i^{\mathsf{V}}\mathbf{W}_i^{\mathsf{O}},\end{aligned} \tag{3}$$

$$\begin{aligned}\mathbf{Z} &= \text{CrossAttention}(\mathbf{Z}, \mathbf{E}) \\ &= \sum_i^h \nu_i \cdot \text{softmax}(\mathbf{Z}\mathbf{W}_i^{\mathsf{Q}'}\mathbf{W}_i^{\mathsf{K}'\top}\mathbf{E}^\top)\mathbf{E}\mathbf{W}_i^{\mathsf{V}'}\mathbf{W}_i^{\mathsf{O}'},\end{aligned} \tag{4}$$

$$\widetilde{\mathbf{X}} = \text{FeedForward}(\mathbf{Z}) = \sum_j^d \xi_j \cdot g(\mathbf{Z}\mathbf{W}_j^1)\mathbf{W}_j^2, \tag{5}$$

where potential bias terms (e.g., linear bias and position bias) are omitted, i means i-th head among h heads, j means j-th intermediate neuron among d neurons, and g is an activation function. We initialize all mask variables to ones to preserve the original structure at the very beginning.

Then expected absolute gradients over either finetuning or pretraining data gives the important scores:

$$\mathbb{I}_i^\mu = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left|\frac{\partial\mathcal{L}(x,y)}{\partial\mu_i}\right|, \tag{6}$$

$$\mathbb{I}_i^\nu = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left|\frac{\partial\mathcal{L}(x,y)}{\partial\nu_i}\right|, \tag{7}$$

$$\mathbb{I}_j^\xi = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left|\frac{\partial\mathcal{L}(x,y)}{\partial\xi_j}\right|, \tag{8}$$

where (x, y) is a data point and $\mathcal{L}$ is the task-specific loss for task-specific models or the language modeling loss for pretrained models. $\mathbb{E}$ represents expectation. The absolute value of gradient for a mask indicates how large the impact of pruning the corresponding structure is, thus implying how important the structure is.

Intuitively, we take a global ranking, in contrast to a local one as in other literature (Hou et al., 2020), for the structures of the same type (i.e., attention head or feed-forward element) from all stacking layers for pruning preference, before which we also normalize the importance scores for same-type structures in a layer with ℓ_2 norm, as suggested by Molchanov et al. (2017), for a balanced pruning. Therefore, for each candidate, we separately prune attention heads and feed-forward elements to the scale so that we reach a qualified structure. For the sake of a corner case that all structures in a module are pruned, we skip the module by feeding the input as the output. While we can alternate to an quite recent pruning method (Xia et al., 2022) exploiting both coarse-grained and fine-grained strategies for state-of-the-art performance, we argue that our framework is agnostic to pruning methods and keep the pruning method simple.

B Dataset Statistics

We conduct experiments on seven datasets. The detailed statistics, maximum sequence lengths, and metrics for datasets we use are shown in Table 8, where the Wikipedia corpus used for pretraining is also attached.

C Additional Implementation Details

The summary of hyperparameters for both task-specific and task-agnostic distillation is shown in Table 9.

D Additional Results upon $\text{BERT}_{\text{base}}$

We further conduct experiments on extremely small scale student model, i.e., $\text{BERT}_{3\%}$. The results are shown in Table 10.

E Practical Inference Measurement

Since FLOPs only offers theoretical inference compute, we additionally provide throughput for empirical inference compute of each model with throughput (i.e., processed tokens per micro second) in Table 11. The test environment is established by feeding 32×128 tokens to models. The amount of decomposed parameters is also attached for a reference.

F Pruned Structure Distribution

We give the distribution of example pruned structures in Figure 4, which exactly show what pruned LMs consist of. While pruned $\text{BERT}_{\text{base}}$ tends to preserve bottom and middle layers, pruned EncT5_{xl}

Table 8: The statistics, maximum sequence lengths, and metrics.

Dataset	#Train exam.	#Dev exam.	Max. length	Metric
SST-2	67K	0.9K	64	Accuracy
MRPC	3.7K	0.4K	128	F1
STS-B	7K	1.5K	128	Spearman Correlation
QQP	364K	40K	128	F1
MNLI-m/mm	393K	20K	128	Accuracy
QNLI	105K	5.5K	128	Accuracy
RTE	2.5K	0.3K	128	Accuracy
Wikipedia	35M	-	128	-

Table 9: The hyperparameters for both task-specific and task-agnostic distillation. The learning rate is searched within different grids for $\text{BERT}_{\text{base}}$ and EncT5_{xl}.

Hyperparameter	Task-specific Distillation	Task-agnostic Distillation
Batch Size	{16,32}	8×128=1024
Optimizer	AdamW	AdamW
Learning Rate	{1e-5, 2e-5, 3e-5}/{1e-4, 2e-4, 3e-4}	3e-4
Training Epochs	10	5
Early-stop Epochs	5	-
Warmup Proportion	0.1	0.01
Weight Decay	0.01	0.01
Sampling Number η	6	3

tends to preserve bottom layers. Meanwhile, neurons in feed-forward layers are more likely to be pruned than heads in attention layers, owing to the centrality of the attention module within an transformer layer.

G Results upon $\text{BERT}_{\text{large}}$

We show extended results of MINIDISC on $\text{BERT}_{\text{large}}$ for readers' interest in Table 12. Consistent patterns have been observed as in $\text{BERT}_{\text{base}}$.

H Results of Small-scale Distillation

When MINIDISC is applied to small $\text{MiniLM}_{12;384\text{H}}$ and $\text{BERT}_{\text{mini}}$ as shown in Table 13, MINIDISC can reversely affect the performance of conventional distillation. Contrarily, MAXIDISC can still improve or at least retain the performance. However, it is less necessary to compress small LMs.

I Additional Task-specific Distillation for TinyBERT

We compare TinyBERT with and without task-specific distillation as in Table 14. The results with task-specific distillation are retrieved from the original paper, since their augmented data is not publicly available. The results demonstrate that TinyBERT is largely supported with task-specific distillation and data augmentation for good performance.

J Negative Derivative-Tradeoff

As mentioned in the main paper, although λ-*tradeoff* is able to provide stable tradeoff measurement, it is dependent on the value of λ. To eliminate this dependency, we design a new measure, negative derivative-tradeoff, which computes the negative derivative of performance to scale at each candidate scale as: $t_a = \lim_{\delta \to 0} \frac{-(m_{a+\delta} - m_a)}{s_{a+\delta} - s_a}$. In the discrete case, $t_{a_i} = \frac{-(m_{a_{i+1}} - m_{a_i})}{\Delta s_a}$. The idea of the measure is basically derived from saving the performance from a potentially significant drop. However, first-order estimation can lead to a high estimation variance and can be further tuned with second-order or so for better performance. The comparison results using λ-*tradeoff* and *ND-tradeoff* are shown in Table 15. It can be seen from the table that MINIDISC-ND also achieves comparable results.

K Varying Schedules for EncT5

Performance variations among possible schedules for EncT5 are displayed in Figure 5, where the existence of scale-performance tradeoff and sufficiency of one teacher assistant can be verified.

Table 10: Additional results of task-specific distillation upon $\text{BERT}_{\text{base}}$.

Method	FLOPs	SST-2	MRPC	STS-B	QQP	MNLI-m/mm	QNLI	RTE	Average
$\mathcal{L}_{\text{TSD3\%}}$	0.3G	85.2	83.6	81.9	82.1	71.9/72.7	81.9	57.4	77.1
$\text{MAXIDISC}_{3\%}$	0.3G	85.6	85.0	82.7	82.7	72.7/72.8	82.0	59.6	77.9
$\text{MINIDISC}_{3\%}$	0.3G	85.9	85.7	83.6	83.1	72.9/73.6	81.9	58.1	78.1

Table 11: Inference compute measurement.

Method	FLOPs	Throughput	Trm params	Emb params
$\text{BERT}_{\text{base}}$	10.9G	55.7tokens/ms	85.7M	23.8M
$\text{BERT}_{10\%}$	1.1G	278.2tokens/ms	9.1M	23.8M
$\text{BERT}_{5\%}$	0.5G	412.9tokens/ms	4.9M	23.8M
$\text{BERT}_{\text{large}}$	38.7G	17.9tokens/ms	303.3M	31.8M
$\text{BERT}_{10\%}$	3.9G	104.1tokens/ms	31.3M	31.8M
$\text{BERT}_{5\%}$	1.9G	154.2tokens/ms	16.3M	31.8M
EncT5_{xl}	155.8G	4.8tokens/ms	1275.1M	32.9M
$\text{EncT5}_{10\%}$	15.6G	38.8tokens/ms	127.4M	32.9M
$\text{EncT5}_{5\%}$	7.8G	64.0tokens/ms	64.0M	32.9M

L Residual Distillation

The results in Table 16 showcase that the follow-up action is at least a no-harm trick.

Table 12: The results of task-specific distillation upon $\text{BERT}_{\text{large}}$.

Method	FLOPs	SST-2	MRPC	STS-B	RTE	Average
$\text{BERT}_{\text{base}}$	10.9G	93.8	91.5	87.1	71.5	86.0
$\mathcal{L}_{\text{TSD10\%}}$	1.1G	88.8	87.8	84.0	66.4	81.8
$\text{MAXIDISC}_{10\%}$	1.1G	89.0	88.2	84.8	66.8	82.2
$\text{MINIDISC}_{10\%}$	1.1G	89.1	88.4	85.4	68.2	82.7
$\mathcal{L}_{\text{TSD5\%}}$	0.5G	85.4	85.5	83.9	63.2	79.5
$\text{MAXIDISC}_{5\%}$	0.5G	86.1	87.0	84.1	65.7	80.7
$\text{MINIDISC}_{5\%}$	0.5G	86.9	87.6	84.8	66.8	81.5
$\text{BERT}_{\text{large}}$	38.7G	94.2	92.5	90.1	75.5	88.1
$\mathcal{L}_{\text{TSD10\%}}$	3.9G	90.4	88.1	87.0	66.1	82.9
$\text{MAXIDISC}_{10\%}$	3.9G	90.6	88.9	87.1	67.2	83.4
$\text{MINIDISC}_{10\%}$	3.9G	90.5	88.8	87.8	66.1	83.3
$\mathcal{L}_{\text{TSD5\%}}$	1.9G	89.2	85.7	85.8	61.4	80.5
$\text{MAXIDISC}_{5\%}$	1.9G	90.4	86.0	85.7	62.8	81.2
$\text{MINIDISC}_{5\%}$	1.9G	89.6	87.4	87.3	61.4	81.4
EncT5_{xl}	155.9G	96.9	95.1	92.3	88.5	93.2
$\mathcal{L}_{\text{TSD10\%}}$	15.6G	94.5	90.2	87.4	67.5	84.9
$\text{MAXIDISC}_{10\%}$	15.6G	94.6	90.5	88.0	70.4	85.9
$\text{MINIDISC}_{10\%}$	15.6G	94.6	91.5	87.8	72.2	86.5
$\mathcal{L}_{\text{TSD5\%}}$	7.8G	92.9	88.0	83.4	58.8	80.8
$\text{MAXIDISC}_{5\%}$	7.8G	93.0	88.0	83.9	67.5	83.1
$\text{MINIDISC}_{5\%}$	7.8G	93.8	89.8	85.3	64.6	83.4

Table 13: The results of task-specific distillation upon small LMs.

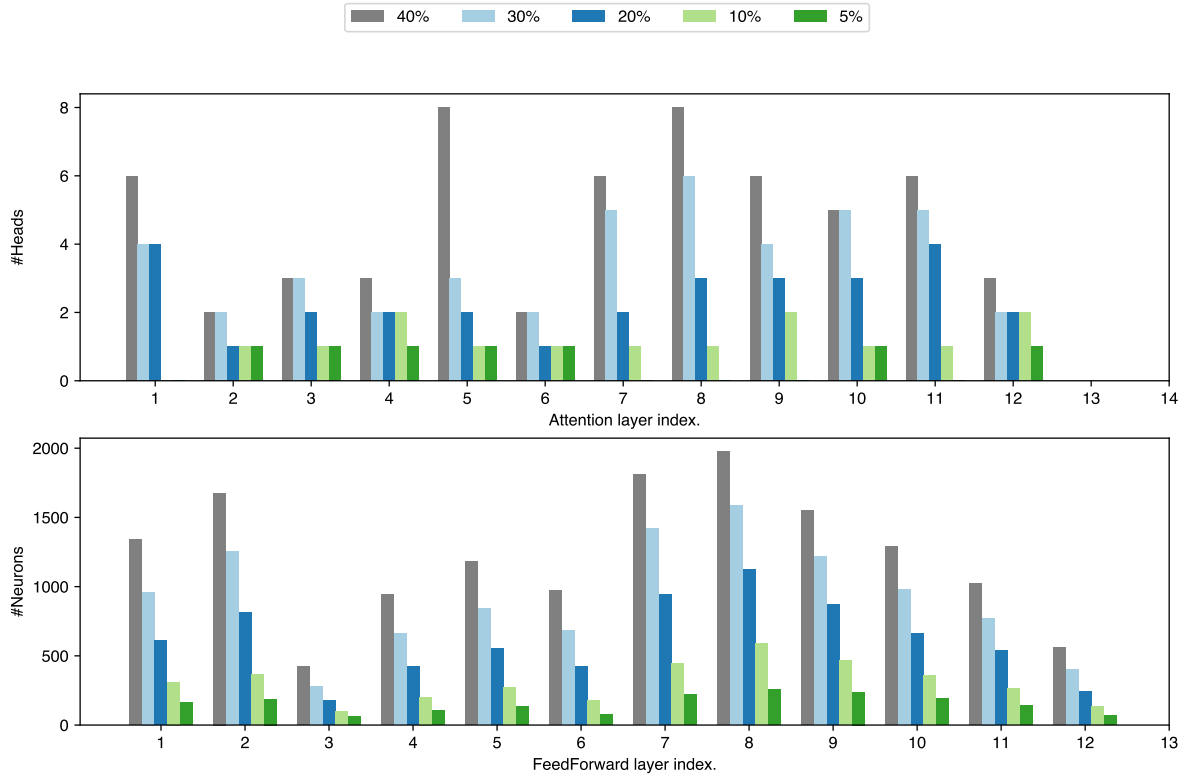
Method	FLOPs	SST-2	MRPC	STS-B	QQP	MNLI-m/mm	QNLI	RTE	Average
$\text{MiniLM}_{\text{12L;384H}}$	2.72G	92.1	90.9	88.6	87.2	83.0/83.3	90.7	72.9	86.1
$\mathcal{L}_{\text{TSD10\%}}$	0.26G	87.8	87.1	85.6	84.3	77.2/78.4	84.8	66.4	81.5
$\text{MAXIDISC}_{10\%}$	0.26G	88.2	88.2	86.3	84.7	77.8/79.2	85.2	65.7	81.9
$\text{MINIDISC}_{10\%}$	0.26G	87.6	86.0	86.5	84.4	77.8/78.6	84.4	64.6	81.3
$\text{BERT}_{\text{mini}}$	0.60G	87.5	86.4	85.3	85.0	76.1/77.2	84.5	66.8	81.1
$\mathcal{L}_{\text{TSD10\%}}$	0.04G	83.3	83.8	81.6	81.6	66.3/71.4	82.7	58.8	76.2
$\text{MAXIDISC}_{10\%}$	0.04G	83.8	84.1	80.7	82.0	66.4/71.6	82.9	58.1	76.2
$\text{MINIDISC}_{10\%}$	0.04G	83.3	82.9	80.6	81.1	67.4/71.3	82.8	58.5	76.0

Table 14: The results of TinyBERT with and without TSD.

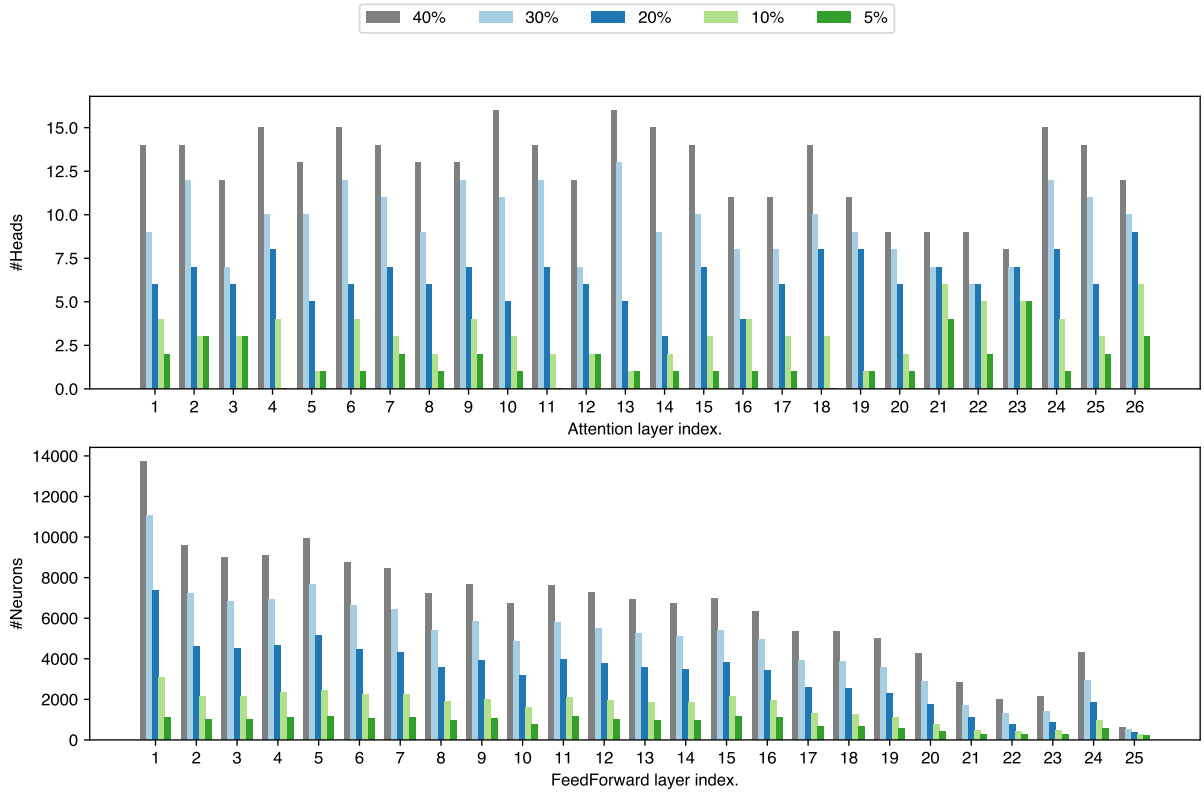
Method	FLOPs	SST-2	MRPC	STS-B	QQP	MNLI-m/mm	QNLI	RTE	Average
$\text{TinyBERT}_{\text{4L;312H}}$ (Jiao et al., 2020)	0.6G	88.5	87.9	86.6	85.6	78.9/79.2	87.3	67.2	82.7
w/ TSD&DA (Jiao et al., 2020)	0.6G	92.7	90.2	86.3	87.1	82.8/82.8	88.0	65.7	84.5
$\text{MiniLM}_{\text{3L;384H}}$ (Wang et al., 2021)	0.7G	89.1	89.1	86.6	85.4	77.8/78.4	87.2	66.1	82.5

Table 15: The results of negative derivative-tradeoff upon $\text{BERT}_{\text{base}}$.

Method	FLOPs	SST-2	MRPC	STS-B	RTE	Average
$\text{BERT}_{\text{base}}$	10.9G	93.8	91.5	87.1	71.5	86.0
$\mathcal{L}_{\text{TSD10\%}}$	1.1G	88.8	87.8	84.0	66.4	81.8
$\text{MAXIDISC}_{10\%}$	1.1G	89.0	88.2	84.8	66.8	82.2
$\text{MINIDISC-}\lambda_{10\%}$	1.1G	89.1	88.4	85.4	68.2	82.7
$\text{MINIDISC-ND}_{10\%}$	1.1G	89.8	87.9	85.4	66.4	82.4
$\mathcal{L}_{\text{TSD5\%}}$	0.5G	85.4	85.5	83.9	63.2	79.5
$\text{MAXIDISC}_{5\%}$	0.5G	86.1	87.0	84.1	65.7	80.7
$\text{MINIDISC-}\lambda_{5\%}$	0.5G	86.9	87.6	84.8	66.8	81.5
$\text{MINIDISC-ND}_{5\%}$	0.5G	86.8	86.0	84.9	66.8	81.1



(a) 12-layer $\text{BERT}_{\text{base}}$.



(b) 24-layer EncT5_{xl}. Layer indices lager than 24 denote modules from the one-layer decoder (i.e., two more attention modules and one more feed-forward modules).

Figure 4: The distribution of example pruned structures. The structures are derived with MRPC dataset.

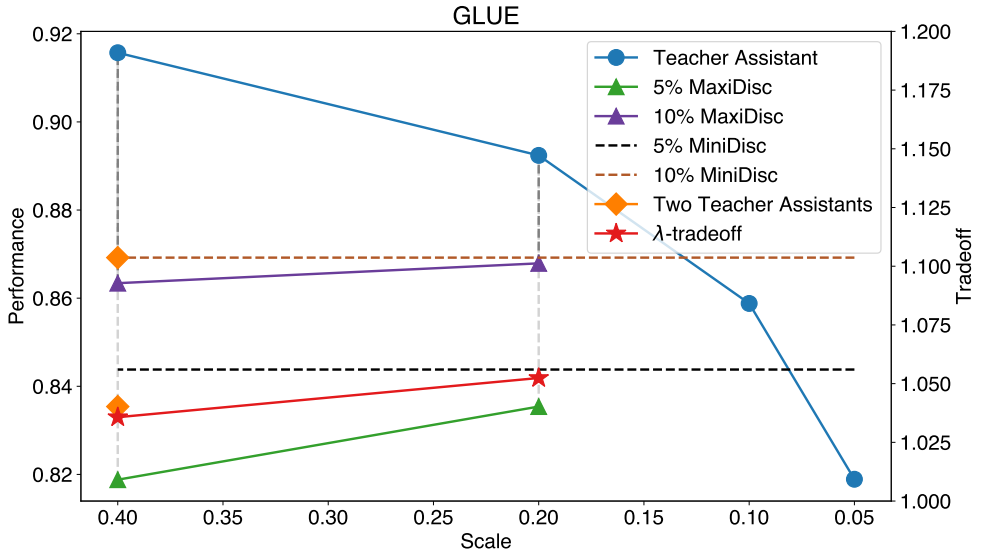


Figure 5: Performance comparisons among various schedules for EncT5. The dots represent performance variations using either one or two teacher assistants for MAXIDISC. The triangles represent performance resulting from MINIDISC using one teacher assistant. The rectangles represent performance resulting from MINIDISC using two teacher assistants.

Table 16: The results of residual distillation upon distilling $\text{BERT}_{\text{base}}$ to $\text{BERT}_{10\%}$.

Method	MRPC	QQP
$\mathcal{L}_{\text{TSD}10\%}$	87.8	84.6
$\text{MINIDISC}_{10\%}$	88.4	84.9
w/ residual distillation	88.4	85.1