One Teacher is Enough? Pre-trained Language Model Distillation from Multiple Teachers

Chuhan Wu[†] **Fangzhao Wu**[‡] **Yongfeng Huang**[†]
[†]Department of Electronic Engineering & BNRist, Tsinghua University, Beijing 100084, China
[‡]Microsoft Research Asia, Beijing 100080, China
{wuchuhan15, wufangzhao}@gmail.com, yfhuang@tsinghua.edu.cn

Abstract

Pre-trained language models (PLMs) achieve great success in NLP. However, their huge model sizes hinder their applications in many practical systems. Knowledge distillation is a popular technique to compress PLMs, which learns a small student model from a large teacher PLM. However, the knowledge learned from a single teacher may be limited and even biased, resulting in low-quality student model. In this paper, we propose a multi-teacher knowledge distillation framework named MT-BERT for pre-trained language model compression, which can train high-quality student model from multiple teacher PLMs. In MT-BERT we design a multi-teacher co-finetuning method to jointly finetune multiple teacher PLMs in downstream tasks with shared pooling and prediction layers to align their output space for better collaborative teaching. In addition, we propose a multi-teacher hidden loss and a multi-teacher distillation loss to transfer the useful knowledge in both hidden states and soft labels from multiple teacher PLMs to the student model. Experiments on three benchmark datasets validate the effectiveness of MT-BERT in compressing PLMs.

1 Introduction

Pre-trained language models (PLMs) such as BERT and RoBERTa have achieved notable success in various NLP tasks (Devlin et al., 2019; Yang et al., 2019; Liu et al., 2019). However, many PLMs have a huge model size and computational complexity, making it difficult to deploy them to low-latency and high-concurrence online systems or devices with limited computational resources (Jiao et al., 2020; Wu et al., 2021).

Knowledge distillation is a widely used technique for compressing large-scale pre-trained language models (Sun et al., 2019; Wang et al., 2020). For example, Sanh et al. (2019) proposed DistilBERT to compress BERT by transferring knowledge from the soft labels predicted by the teacher model to student model with a distillation loss. Jiao et al. (2020) proposed TinyBERT, which aligns the hidden states and the attention heatmaps between student and teacher models. These methods usually learn the student model from a single teacher model (Gou et al., 2020). However, the knowledge and supervision provided by a single teacher model may be insufficient to learn an accurate student model, and the student model may also inherit the bias in the teacher model (Bhardwaj et al., 2020). Fortunately, many different large PLMs such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and UniLM (Dong et al., 2019) are off-the-shelf. These PLMs may encode complementary knowledge because they usually have different configurations and are trained on different corpus with different self-supervision tasks (Qiu et al., 2020). Thus, incorporating multiple pre-trained language models into knowledge distillation has the potential to learn better student models.

In this paper, we present a multi-teacher knowledge distillation method named MT-BERT for pre-trained language model compression.[1] In MT-BERT, we propose a multi-teacher co-finetuning framework to jointly finetune multiple teacher models with a shared pooling and prediction module to align their output hidden states for better collaborative student teaching. In addition, we propose a multi-teacher hidden loss and a multi-teacher distillation loss to transfer the useful knowledge in both hidden states and soft labels from multiple teacher models to student model. Experiments on three benchmark datasets show MT-BERT can effectively improve the quality of student models for PLM compression and outperform many single-teacher knowledge distillation methods.

[1]We focus on task-specific knowledge distillation.

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 4408–4413

2 MT-BERT

Next, we introduce the details of our multi-teacher knowledge distillation method MT-BERT for pre-trained language model compression.[2] We first introduce the multi-teacher co-finetuning framework to jointly finetune multiple teacher models in downstream tasks, and then introduce the multi-teacher distillation framework to collaboratively teach the student with multiple teachers.

2.1 Multi-Teacher Co-Finetuning

Researchers have found that distilling the knowledge in the hidden states of a teacher model is important for effective student teaching (Sun et al., 2019; Jiao et al., 2020). However, since different teacher PLMs are separately pre-trained with different settings, finetuning them independently may lead to some inconsistency in their feature space, which is not optimal for transferring knowledge in the hidden states of multiple teachers. Thus, we design a multi-teacher co-finetuning framework to obtain some uniformity among the hidden states output by the last layer of different teacher models for better collaborative student teaching, as shown in Fig. 1. Assume there are N teacher models, and denote the hidden states output by the top layer of the i-th teacher as $\mathbf{H}^i$. We use a shared pooling[3] layer to summarize each hidden matrix $\mathbf{H}^i$ into a unified text embedding, and then use a shared dense layer to convert it into a soft probability vector $\mathbf{y}_i$. Finally, we jointly optimize the summation of the task-specific losses of all teacher models, i.e., $\sum_{i=1}^{N} \mathrm{CE}(\mathbf{y}, \mathbf{y}_\mathrm{i})$, where $\mathrm{CE}(\cdot,\cdot)$ stands for the cross-entropy loss and $\mathbf{y}$ is the ground-truth label. Since the pooling and prediction layers are shared among different teachers, the feature space of the output hidden states from different teacher PLMs can be aligned, which can help them collaborate better for student teaching.

2.2 Multi-Teacher Knowledge Distillation

Next, we introduce our proposed multi-teacher knowledge distillation framework, which is shown in Fig. 2. Two loss functions are used for knowledge distillation, i.e., a multi-teacher hidden loss and a multi-teacher distillation loss.

The multi-teacher hidden loss aims to transfer knowledge in the hidden states of multiple teachers.

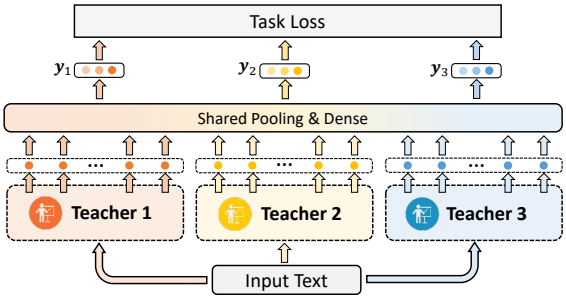


Figure 1: The multi-teacher co-finetuning framework.

Assume there are N teacher PLMs, and each of them has $T \times K$ Transformer layers. They collaboratively teach a student model with K layers, and each layer in the student model corresponds to T layers in teacher PLMs.[4] Denote the hidden states output by the j-th layer of the student model as $\mathbf{H}^s_j$, and the corresponding hidden states output by the $(T \times j)$-th layer of the i-th teacher model as $\mathbf{H}^i_{Tj}$. Following (Sun et al., 2019), we apply the mean squared error (MSE) to the hidden states of corresponding layers in the student and teacher models to encourage the student model to have similar functions with teacher models. The multi-teacher hidden loss $\mathcal{L}_{MT-Hid}$ is formulated as follows:

$$\mathcal{L}_{MT-Hid} = \sum_{i=1}^{N}\sum_{j=1}^{T} \mathrm{MSE}(\mathbf{H}^{\mathrm{s}}_{\mathrm{j}}, \mathbf{W}_{\mathrm{ij}}\mathbf{H}^{\mathrm{i}}_{\mathrm{Tj}}), \quad (1)$$

where $\mathbf{W}_{ij}$ is a learnable transformation matrix.

The multi-teacher distillation loss aims to transfer the knowledge in the soft labels output by multiple teachers to student. The predictions of different teachers on the same sample may have different correctness and confidence. Thus, it may be suboptimal to simply ensemble (Fukuda et al., 2017; Liu et al., 2020) or choose (Yuan et al., 2020) soft labels without the help of task labels. Since in task-specific knowledge distillation the labels of training samples are available, we propose a distillation loss weighting method to assign different weights to different samples. The weights are based on the loss inferred from the predictions of corresponding teacher against the gold labels. More specifically, the multi-teacher distillation loss $\mathcal{L}_{MT-Dis}$ is formulated as follows:

$$\mathcal{L}_{MT-Dis} = \sum_{i=1}^{N} \frac{\mathrm{CE}(\mathbf{y}_\mathrm{i}/\mathrm{t}, \mathbf{y}_\mathrm{s}/\mathrm{t})}{1 + \mathrm{CE}(\mathbf{y}, \mathbf{y}_\mathrm{i})}, \quad (2)$$

[2] Codes available at https://github.com/wuch15/MT-BERT

[3] In MT-BERT we use attentive pooling because it performs better than average pooling and "[CLS]" token embedding.

[4] Here we assume that all teacher models have the same number of layers. We will explore to generalize MT-BERT to scenarios where teacher models have different architectures in our future work.

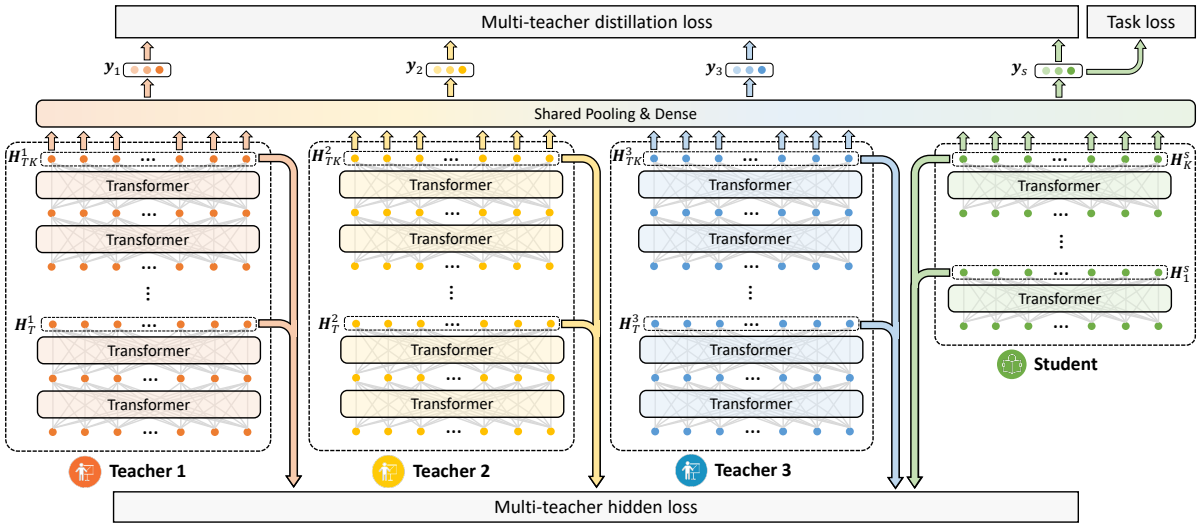


Figure 2: The multi-teacher knowledge distillation framework in MT-BERT.

where t is the temperature coefficient. In this way, if a teacher's prediction on a certain sample is more close to the ground-truth label, its corresponding distillation loss will gain higher weight.

Following (Tang et al., 2019; Lu et al., 2020), we also incorporate gold labels to compute the task-specific loss $\mathcal{L}_{Task}$ based on the predictions of the student model, i.e., $\mathcal{L}_{Task} = \mathrm{CE}(\mathbf{y}, \mathbf{y}_s)$. The final loss function $\mathcal{L}$ for learning the student model is a summation of the multi-teacher hidden loss, multi-teacher distillation loss and the task-specific loss, which is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{MT-Hid} + \mathcal{L}_{MT-Dis} + \mathcal{L}_{Task}. \quad (3)$$

3 Experiments

3.1 Datasets and Experimental Settings

We conduct experiments on three benchmark datasets with different sizes. The first one is SST-2 (Socher et al., 2013), which is a benchmark for text sentiment classification. The second one is RTE (Bentivogli et al., 2009), which is a widely used dataset for natural language inference. The third one is the MIND dataset (Wu et al., 2020c), which is a large-scale public English news dataset.[5] We perform the news topic classification task on this dataset. The detailed statistics of the three datasets are shown in Table 1.

In our experiments, we use the pre-trained 12-layer BERT, RoBERTa and UniLM (Bao et al., 2020)[6] models as the teachers to distill a 6-layer and a 4-layer student models respectively. We use the token embeddings and the first 4 or 6 Transformer layers of UniLM to initialize the parameters of the student model. The pooling layer is implemented by an attention network (Yang et al., 2016; Wu et al., 2020a). The temperature coefficient t is set to 1. The attention query dimension in the attentive pooling layer is 200. The optimizer we use is Adam (Bengio and LeCun, 2015). The teacher model learning rate is 2e-6 while the student model learning rate is 5e-6. The batch size is 64. Following (Jiao et al., 2020), we report the accuracy score on the SST-2 and RTE datasets. In addition, since the news topics in the MIND dataset are highly imbalanced, following (Wu et al., 2020b) we report both accuracy and macro-F1 scores. Each experiment is independently repeated 5 times and the average scores are reported.

Dataset	#Train	#Dev	#Test	#Class
SST-2	67k	872	1.8k	2
RTE	2.5k	276	3.0k	2
MIND	102k	2.6k	26k	18

Table 1: The statistics of the three datasets.

3.2 Performance Evaluation

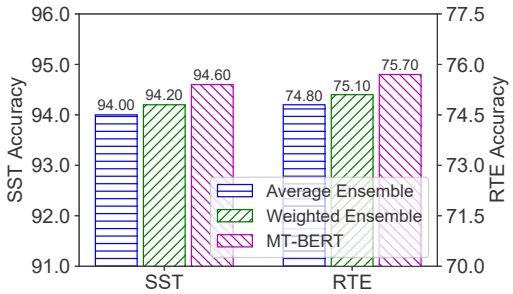
We compare the performance of MT-BERT with two groups of baselines. The first group includes the 12-layer version of the teacher models, i.e., BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and UniLM (Bao et al., 2020). The second group includes the 6-layer and 4-layer student

[5] https://msnews.github.io/

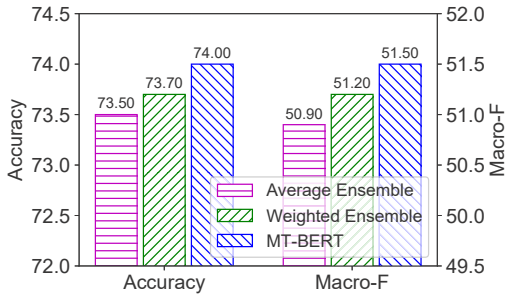
[6] We used the UniLMv2 version.

Methods	SST-2 (Acc.)	RTE (Acc.)	MIND (Acc./Macro-F)		#Param
$BERT_{12}$	92.8	68.6	73.6	51.3	109M
$RoBERTa_{12}$	94.8	78.7	73.9	51.5	109M
$UniLM_{12}$	**95.1**	**81.3**	**74.6**	**51.9**	109M
$DistilBERT_6$	92.5	58.4	72.5	50.4	67.0M
$DistilBERT_4$	91.4	54.1	72.1	50.2	52.2M
$BERT\text{-}PKD_6$	92.0	65.5	72.7	50.6	67.0M
$BERT\text{-}PKD_4$	89.4	62.3	72.4	50.3	52.2M
$TinyBERT_6$	93.1	70.0	73.4	50.8	67.0M
$TinyBERT_4$	92.6	66.6	73.0	50.4	14.5M
$MT\text{-}BERT_6$	**94.6**	**75.7**	**74.0**	**51.5**	67.0M
$MT\text{-}BERT_4$	93.9	73.8	73.8	51.2	52.2M

Table 2: Results and parameters of different methods.



(a) SST-2 and RTE datasets.



(b) MIND dataset.

Figure 3: Comparison of MT-BERT and ensemble-based multi-teacher distillation methods.

models distilled by DistilBERT (Sanh et al., 2019), BERT-PKD (Sun et al., 2019) and TinyBERT (Jiao et al., 2020), respectively. The results of different methods are summarized in Table 2.[7] Referring to this table, we find MT-BERT can consistently outperform all the single-teacher knowledge distillation methods compared here. This is because the knowledge provided by a single teacher model may be insufficient, and incorporating the complementary knowledge encoded in multiple teacher models can help learn better student model. In addition,

[7]We take the original reported results of baseline methods on the SST-2 and RTE datasets, and we run their codes to obtain their results on the *MIND* dataset.

Teachers	SST-2 (Acc.)	RTE (Acc.)	MIND (Acc./Macro-F)	
BERT	92.1	65.8	72.8	50.6
RoBERTa	92.9	68.9	73.0	50.7
UniLM	93.3	70.6	73.4	50.9
BERT+RoBERTa	93.6	71.2	73.3	50.9
BERT+UniLM	93.9	73.7	73.6	51.1
RoBERTa+UniLM	94.3	74.9	73.7	51.3
All	94.6	75.7	74.0	51.5

Table 3: Different combinations of teacher models.

compared with the teacher models, MT-BERT has much fewer parameters and its performance is comparable or even better than these teacher models. It shows that MT-BERT can effectively inherit the knowledge of multiple teacher models even if the model size is significantly compressed.

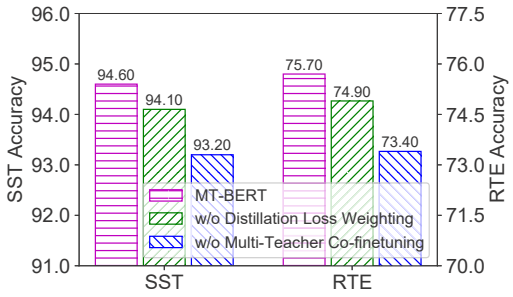
We also compare MT-BERT with several multi-teacher knowledge distillation methods proposed in the computer vision field that ensemble the outputs of different teachers for student teaching (You et al., 2017; Liu et al., 2020). The results are shown in Fig. 3. We find our MT-BERT performs better than these ensemble-based multi-teacher knowledge distillation methods. This is because these methods do not consider the correctness of the teacher model predictions on a specific sample and cannot transfer useful knowledge encoded in the intermediate layers, which may not be optimal for collaborative knowledge distillation from multiple teachers.

3.3 Effectiveness of Multiple Teachers

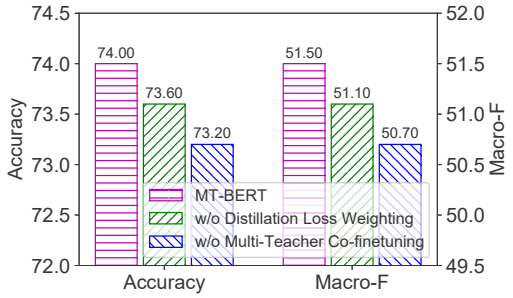
Next, we study the effectiveness of using multiple teacher PLMs for knowledge distillation. We compare the performance of the 6-layer student model distilled from different combinations of teacher models. The results are summarized in Table 3. It shows that using multiple teacher PLMs can achieve better performance than using a single one. This is because different teacher models can encode complementary knowledge and combining them together can provide better supervision for student model. In addition, combining all three teacher PLMs can further improve the performance of student model, which validates the effectiveness of MT-BERT in distilling knowledge from multiple teacher models.

3.4 Ablation Study

We study the effectiveness of the two important techniques in MT-BERT, i.e., the multi-teacher co-finetuning framework and the distillation loss

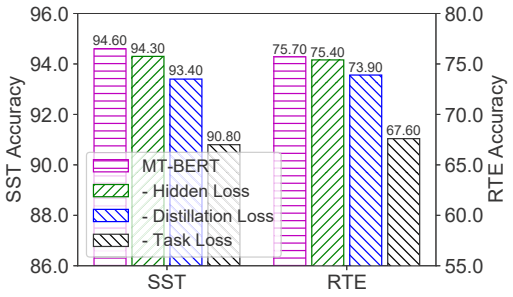


(a) SST-2 and RTE datasets.

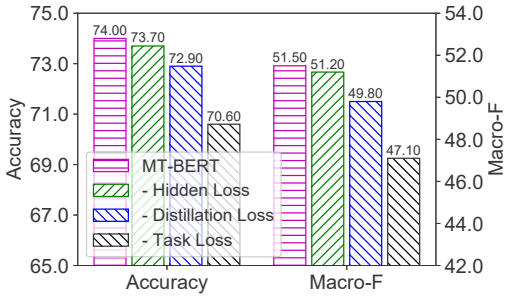


(b) MIND dataset.

Figure 4: Effectiveness of multi-teacher co-finetuning and distillation loss weighting.



(a) SST-2 and RTE datasets.



(b) MIND dataset.

Figure 5: Effectiveness of different loss functions.

weighting method. We compare MT-BERT and its variants with one of these modules removed, as shown in Fig. 4. The student model has 6 layers. We find the multi-teacher co-finetuning framework is very important. This is because the hidden states of different teacher models can be in very different spaces, and jointly finetuning multiple teachers with shared pooling and prediction layers can align their output hidden spaces for better collaborative student teaching. In addition, the distillation loss weighting method is also useful. This is because the predictions of different teachers on the same sample may have different correctness, and focusing on the more reliable predictions is helpful for distilling accurate student models.

We also verify the effectiveness of different loss functions in MT-BERT, which is shown in Fig. 5. We find the task loss is very important. It is because in our experiments the corpus for task-specific distillation are not large and the direct supervision from task labels is useful. In addition, the distillation loss is also important. It indicates that transferring the knowledge in soft labels plays a critical role in knowledge distillation. Moreover, the hidden loss is also helpful. It shows that hidden states of different teacher models can provide useful knowledge for student model learning.

4 Conclusion

In this paper, we propose a multi-teacher knowledge distillation method named MT-BERT for pre-trained language model compression, which can learn small but strong student model from multiple teacher PLMs in a collaborative way. We propose a multi-teacher co-finetuning framework to align the output hidden states of multiple teacher models for better collaborative student teaching. In addition, we design a multi-teacher hidden loss and a multi-teacher distillation loss to transfer the useful knowledge in both hidden states and prediction of multiple teacher models to student model. The extensive experiments on three benchmark datasets show that MT-BERT can effectively improve the performance of pre-trained language model compression, and can outperform many single-teacher knowledge distillation methods.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant numbers U1936208, U1936216, U1836204, and U1705261. We thank Xing Xie, Tao Qi, Ruixuan Liu and Tao Di for their great comments and suggestions which are important for improving this work.

References

Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, et al. 2020. Unilmv2: Pseudo-masked language models for unified language model pre-training. In *ICML*, pages 642–652. PMLR.

Yoshua Bengio and Yann LeCun. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The fifth pascal recognizing textual entailment challenge. In *TAC*.

Rishabh Bhardwaj, Navonil Majumder, and Soujanya Poria. 2020. Investigating gender bias in BERT. *arXiv preprint arXiv:2009.05021*.

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.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *NeurIPS*, pages 13042–13054.

Takashi Fukuda, Masayuki Suzuki, Gakuto Kurata, Samuel Thomas, Jia Cui, and Bhuvana Ramabhadran. 2017. Efficient knowledge distillation from an ensemble of teachers. In *Interspeech*, pages 3697–3701.

Jianping Gou, Baosheng Yu, Stephen John Maybank, and Dacheng Tao. 2020. Knowledge distillation: A survey. *arXiv preprint arXiv:2006.05525*.

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 Findings*, pages 4163–4174.

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 preprint arXiv:1907.11692*.

Yuang Liu, Wei Zhang, and Jun Wang. 2020. Adaptive multi-teacher multi-level knowledge distillation. *Neurocomputing*, 415:106–113.

Wenhao Lu, Jian Jiao, and Ruofei Zhang. 2020. Twinbert: Distilling knowledge to twin-structured compressed bert models for large-scale retrieval. In *CIKM*, pages 2645–2652.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, pages 1–26.

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

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.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019. Patient knowledge distillation for bert model compression. In *EMNLP-IJCNLP*, pages 4314–4323.

Raphael Tang, Yao Lu, Linqing Liu, Lili Mou, Olga Vechtomova, and Jimmy Lin. 2019. Distilling task-specific knowledge from bert into simple neural networks. *arXiv preprint arXiv:1903.12136*.

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*.

Chuhan Wu, Fangzhao Wu, Tao Qi, Xiaohui Cui, and Yongfeng Huang. 2020a. Attentive pooling with learnable norms for text representation. In *ACL*, pages 2961–2970.

Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2020b. Improving attention mechanism with query-value interaction. *arXiv preprint arXiv:2010.03766*.

Chuhan Wu, Fangzhao Wu, Yang Yu, Tao Qi, Yongfeng Huang, and Qi Liu. 2021. Newsbert: Distilling pre-trained language model for intelligent news application. *arXiv preprint arXiv:2102.04887*.

Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, et al. 2020c. Mind: A large-scale dataset for news recommendation. In *ACL*, pages 3597–3606.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*, pages 5753–5763.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *NAACL-HLT*, pages 1480–1489.

Shan You, Chang Xu, Chao Xu, and Dacheng Tao. 2017. Learning from multiple teacher networks. In *KDD*, pages 1285–1294.

Fei Yuan, Linjun Shou, Jian Pei, Wutao Lin, Ming Gong, Yan Fu, and Daxin Jiang. 2020. Reinforced multi-teacher selection for knowledge distillation. *arXiv preprint arXiv:2012.06048*.