ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096045

BOOSTING PROMPT-BASED FEW-SHOT LEARNERS THROUGH OUT-OF-DOMAIN KNOWLEDGE DISTILLATION

Xiaoqing Chen[1], Chengyu Wang[2], Junwei Dong[1], Minghui Qiu[2,†], Liang Feng[1,†], Jun Huang[2]

[1] College of Computer Science, Chongqing University, Chongqing, China
[2] Alibaba Group, Hangzhou, China

ABSTRACT

Prompt-based learning improves the performance of Pre-trained Language Models (PLMs) over few-shot learning and is suitable for low-resourced scenarios. However, it is challenging to deploy large PLMs online. Knowledge Distillation (KD) can compress large PLMs into small ones; yet, few-shot KD for prompt-tuned PLMs is challenging due to the lack of training data and the capacity gap between teacher and student models. We propose Boost-Distiller, the first few-shot KD algorithm for prompt-tuned PLMs with the help of the out-of-domain data. Apart from distilling the model logits, Boost-Distiller specifically considers heuristically-generated fake logits that improve the generalization abilities of student models. We further leverage the cross-domain model logits, weighted with domain expertise scores that measure the transferablity of out-of-domain instances. Experiments over various datasets show Boost-Distiller consistently outperforms baselines by a large margin.

***Index Terms*—** knowledge distillation, few-shot learning, transfer learning, pre-trained language model

1. INTRODUCTION

Pre-trained Language Models (PLMs) have greatly boosted the performance of various NLP tasks based on the "pre-training and fine-tuning" framework [1]. Yet, the performance of PLMs is still limited by the number of labeled training samples. Recently, prompt-based learning is proposed to reformulate NLP tasks as cloze questions and to provide additional task guidance by discrete or continuous prompts [2, 3, 4, 5, 6], which further enables effective few-shot learning for PLMs and especially suitable for low-resourced scenarios.

The "secret ingredient" in prompt-based learning is that they directly leverage rich pre-training knowledge in PLMs as the "prior knowledge" for downstream tasks, such as Masked Language Modeling (MLM) for BERT-style models [7]. Hence, large-scale PLMs typically have better few-shot performance due to large model capacity, which unfortunately makes it challenging to deploy them in resource-constrained environments.

Knowledge Distillation (KD) aims to compress a large model into a small one while keeping the model performance as much as possible [8]. Although there are a variety of works focusing on KD for PLMs (such as [9, 10, 11]), distilling prompt-based few-shot learners in low-resourced scenarios is non-trivial for several reasons: i) existing KD algorithms for BERT-style models (mentioned above) are not designed for prompt-based PLMs; ii) the lack of training instances makes supervised signals highly insufficient when the knowledge is transferred from teacher to student models; and iii) there exist large capacity gaps between PLMs of different sizes [12, 13], often resulting in weak students during KD. Hence, a natural question arises: how can we effectively distill prompt-based few-shot learners to smaller models with few training instances?

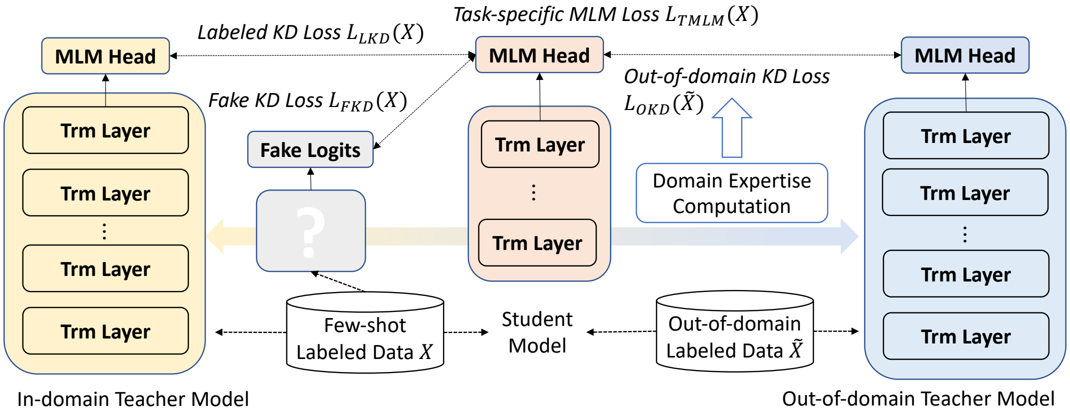
In this paper, we propose Boost-Distiller, the first few-shot KD algorithm for prompt-based learners, with the help of out-of-domain datasets. As teacher and student models have significantly different capacities, in contrast to previous works, we observe that distilling intermediate-layer representations for few-shot learning may have clear negative impacts on the KD performance. Thus, only the logits of the MLM heads are employed as the knowledge signals. To address the data-hungry issue, we further consider distilling i) the MLM logits of the out-of-domain model, weighted with domain expertise scores measuring the transferablity of out-of-domain instances and ii) the heuristically-generated fake logits that improve the generalization abilities of student models. In the experiments, we evaluate Boost-Distiller over eight public NLP datasets. Experiments show that it consistently outperforms baselines by a large margin.

2. BOOST-DISTILLER: THE PROPOSED APPROACH

The Boost-Distiller framework is presented in Figure 1. Given an N-way-K-shot training set $X = \{(x_i, y_i)\}$ (where $y_i \in \mathcal{Y}$ is the label of the text x_i with the class label set $|\mathcal{Y}| = N$) and a prompt-tuned teacher model parameterized by Θ_T, the goal is to obtain a smaller prompt-tuned PLM

X. Chen and C. Wang contributed equally to this work.

[†] Corresponding authors: L. Feng (liangf@cqu.edu.cn) and M. Qiu (minghui.qmh@alibaba-inc.com).



Labeled KD Loss $L_{LKD}(X)$
Task-specific MLM Loss $L_{TMLM}(X)$
MLM Head
MLM Head
MLM Head
Fake KD Loss $L_{FKD}(X)$
Out-of-domain KD Loss $L_{OKD}(\tilde{X})$
Fake Logits
?
Trm Layer
Trm Layer
Trm Layer
Trm Layer
Trm Layer
Trm Layer
Domain Expertise Computation
Trm Layer
Trm Layer
Trm Layer
Trm Layer
Few-shot Labeled Data X
Student Model
Out-of-domain Labeled Data $\tilde{X}$
In-domain Teacher Model
Out-of-domain Teacher Model

Fig. 1. An illustration of the Boost-Distiller framework.

parameterized by Θ_S that preserves the similar performance compared to that of Θ_T. As the size of X is extremely small (i.e., $N \times K$), the supervised signals for KD are insufficient. We also assume that there is a larger, out-of-domain dataset $\tilde{X} = \{(x_i, y_i)\}$ with $|\tilde{X}| = n \cdot |X|$, serving as the auxiliary dataset for KD. Without loss of generality, we use PET [3] to prompt-tune the teacher and student models throughout our work to ensure good few-shot performance.

2.1. Learning from In-domain Teacher

As described previously, the prediction results of our PLMs are generated by the MLM head. Following PET [3], let $l(y)$ be the label word for the class y, and $s_{\Theta_T}(l(y)|x_i)$ be the score of predicting $l(y)$ at the MLM token w.r.t. the input x_i and the PLM Θ_T. The probability of x_i being assigned to the class y is:

$$p_T(y|x_i) = \frac{\exp\{s_{\Theta_T}(l(y)|x_i)\}}{\sum_{y' \in \mathcal{Y}} \exp\{s_{\Theta_T}(l(y')|x_i)\}}. \quad (1)$$

Denote $p_T(\vec{y}|x_i)$ as the probability vector across all N classes $\mathcal{Y}$ where $\vec{y}_i$ is the one-hot ground-truth vector for x_i. The task-specific classification loss for the student model is defined as follows:

$$\mathcal{L}_{\text{TMLM}}(X) = \frac{1}{|X|} \sum_{(x_i, y_i) \in X} \text{CE}(\vec{y}_i, p_S(\vec{y}|x_i)) \quad (2)$$

where CE$(\cdot, \cdot)$ denotes the cross-entropy loss between the two vectors, and $p_S(\vec{y}|x_i)$ is the result generated from the student model.

During KD, we wish the student model to learn from its prompt-tuned teacher model. However, due to capacity differences, in exploratory experiments (which will be presented), we add the loss function for various elements proposed in [11] and find that distilling too many elements from the teacher model has negative impacts. Here, we define the labeled KD loss purely based on the logits of the MLM heads, shown as follows:

$$\mathcal{L}_{\text{LKD}}(X) = \frac{1}{|X|} \sum_{(x_i, y_i) \in X} \text{CE}\left(\frac{p_T(\vec{y}|x_i)}{\alpha}, p_S(\vec{y}|x_i)\right) \quad (3)$$

where $\alpha > 0$ is the temperature factor. This loss is proved to be very useful in our task scenario.

Despite $\mathcal{L}_{\text{LKD}}(X)$, the lack of training data still makes supervised signals rather limited for KD. Inspired by [14], we mimic the behavior of the teacher model and generate fake logits for the student to learn. Specifically, we derive the fake probability distribution $p_F(\vec{y}|x_i)$ based on the label smoothing operation where

$$p_F(y|x_i) = \begin{cases} M & (y = y_i) \\ \frac{1-M}{N-1} & (y \neq y_i) \end{cases} \quad (4)$$

with M to be a constant close to 1. The fake logits are generated by converting $p_F(\vec{y}|x_i)$ to the logits vector $l_F(\vec{y}|x_i)$ by setting a high temperature (as suggested by [14]). Thus, the fake KD loss $\mathcal{L}_{\text{FKD}}(X)$ is defined as follows:

$$\mathcal{L}_{\text{FKD}}(X) = \frac{1}{|X|} \sum_{(x_i, y_i) \in X} \text{CEL}(\vec{y}_i, l_F(\vec{y}|x_i))) \quad (5)$$

where CEL$(\cdot, \cdot)$ is the cross-entropy loss with logits, defined between two vectors.

2.2. Learning from Out-of-domain Teacher

In practice, applying $\mathcal{L}_{\text{FKD}}(X)$ alone is still insufficient to alleviate the data-hungry problem as it only provides $N \times K$ additional signals. We further leverage a non-few-shot out-of-domain dataset $\tilde{X} = \{(x_i, y_i)\}$ for KD, which is relatively easier to obtain in low-resourced situations. A naive approach for cross-domain KD is to apply $\mathcal{L}_{\text{LKD}}(\tilde{X})$ over the dataset $\tilde{X}$. Yet, the domain gap between X and $\tilde{X}$ causes the student model to capture the non-transferable knowledge from

Paradigm	Method	MNLI	MNLI-mm	SNLI	SST-2	MR	MRPC	QQP	QNLI	RTE	Avg.
FT	*Teacher FT (Upper Bound)*	*44.3*	*46.3*	*48.5*	*79.7*	*75.0*	*71.1*	*60.4*	*60.7*	*50.5*	*59.6*
	Student FT (Lower Bound)	34.8	35.1	41.4	63.5	51.3	57.1	55.0	60.6	51.3	50.0
	Vanilla KD [9]	34.9	35.6	40.6	52.5	50.2	56.6	45.8	56.7	50.2	47.0
	TinyBERT [11]	40.6	41.5	42.8	63.5	53.3	65.4	58.2	57.2	53.3	52.9
PT	*Teacher PT (Upper Bound)*	*67.0*	*69.0*	*77.3*	*93.1*	*84.5*	*66.4*	*71.9*	*65.5*	*71.3*	*74.3*
	Student PT (Lower Bound)	35.4	36.2	38.9	60.2	55.8	61.0	55.4	54.5	51.8	49.9
	Prompt-KD	35.5	36.0	39.3	61.9	53.7	63.5	53.4	53.6	51.6	49.8
PT	**Boost-Distiller (Ours)**	**45.6** (SNLI)	**47.6** (SNLI)	**47.3** (MNLI)	**76.2** (MR)	**73.6** (SST-2)	**67.9** (QQP)	**60.8** (MRPC)	**60.2** (RTE)	**53.8** (QNLI)	**59.2**

Table 1. Comparison between the proposed Boost-Distiller and baselines in terms of accuracy for few-shot KD (%). Dataset names in brackets refer to the corresponding out-of-domain datasets. "FT" and "PT" refer to traditional fine-tuning and prompt-tuning, respectively.

$\tilde{X}$, thus lowering the model performance. This problem becomes particularly severe when $|\tilde{X}| >> |X|$ (which is exactly the case in our work).

In Boost-Distiller, we propose the domain expertise score that effectively measures whether an out-of-domain instance $(x_i, y_i) \in \tilde{X}$ is useful for KD without human labeling. To ensure model homogeneity, we also train a PET-based teacher model over $\tilde{X}$, parameterized by Θ_{OT}. The instance (x_i, y_i) is passed to both Θ_{OT} and Θ_T to obtain the prediction results $p_{OT}(\vec{y}|x_i)$ and $p_T(\vec{y}|x_i)$. The score s_i is computed based on the Jensen-Shannon Divergence (JSD) between the two probability vectors w.r.t. the instance (x_i, y_i), i.e.,

$$\begin{aligned} s_i =& \frac{1}{2}(\text{KLD}(p_{OT}(\vec{y}|x_i)||p_T(\vec{y}|x_i)) \\ &+ \text{KLD}(p_T(\vec{y}|x_i)||p_{OT}(\vec{y}|x_i))) \end{aligned} \tag{6}$$

where KLD$(\cdot||\cdot)$ is the Kullback–Leibler Divergence (KLD) between two probabilistic distributions. Based on the domain expertise, we then define the out-of-domain KD loss $\mathcal{L}_{\text{OKD}}(\tilde{X})$ as follows:

$$\mathcal{L}_{\text{OKD}}(\tilde{X}) = \frac{\sum_{(x_i,y_i)\in\tilde{X}} s_i \cdot \text{CE}\left(\frac{p_{OT}(\vec{y}|x_i)}{\alpha}, p_S(\vec{y}|x_i)\right)}{|\tilde{X}|}. \tag{7}$$

We have also tested other techniques such as employing the fake logits for out-of-domain KD. However, the performance improvement remains minimal, mostly because using the non-few-shot dataset $\tilde{X}$ alone provides sufficient signals for cross-domain knowledge transfer.

In summary, the overall loss function $\mathcal{L}(X, \tilde{X})$ of our Boost-Distiller framework is:

$$\begin{aligned} \mathcal{L}(X, \tilde{X}) =& \mathcal{L}_{\text{TMLM}}(X) + \lambda_1 \mathcal{L}_{\text{LKD}}(X) \\ &+ \lambda_1 \mathcal{L}_{\text{FKD}}(X) + \lambda_2 \mathcal{L}_{\text{OKD}}(\tilde{X}) \end{aligned} \tag{8}$$

where λ_1 and λ_2 are balancing hyper-parameters.

3. EXPERIMENTS

In this section, we conduct extensive experiments to evaluate Boost-Distiller on various aspects.

3.1. Datasets and Experimental Settings

In the experiments, we employ eight public datasets to evaluate the Boost-Distiller framework, which are divided into three groups: natural language inference (MNLI [15], SNLI [16], QNLI [17] and RTE [18]), question answering (MRPC [19] and QQP[1]) and sentiment analysis (MR [20] and SST-2 [21]).

We use RoBERTa-large [22] as the teacher model (with around 355M parameters) and BERT-small [23] as the student model. Following [4], we have $K = 16$ and test our model over five different few-shot training sets. For out-of-domain data, we have $n = 10$ and also vary the number of out-of-domain data instances in detailed analysis. We keep all prompts to be the same as PET [3]. During training, we fix the batch size and the learning rate to be 4 and 1e-5, respectively. Other hyper-parameters (λ_1, λ_2 and α) are tuned on development sets. For evaluation, we report the average model performance in terms of accuracy (with the same random seeds for all methods). We implement Boost-Distiller in PyTorch and conduct experiments on Tesla V100 GPUs.

3.2. Main Results

The results are shown in Table 1. Two paradigms for tuning PLMs are used for comparison, namely standard fine-tuning (FT) and prompt-tuning (PT). For each paradigm, we also list the performance of the respective teacher and student models as upper and lower bounds. The baselines include vanilla KD [9] and TinyBERT [11] for FT, and Prompt-KD for PT (which distills the logits of the teacher MLM head only). Based on the results, we draw the following conclusions. i) Due to the lack of labeled training data, KD ap-

[1] https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs

Table 2. Ablation study of Boost-Distiller (%).

Method	SST-2	MR	QNLI	RTE
Full Implement.	**76.2**	**73.6**	**60.2**	**53.8**
w/o. teacher logits	74.9	73.4	58.8	53.0
w/o. fake logits	74.9	73.4	59.2	52.4
w/o. domain expertise	72.4	73.0	58.1	49.8
w/o. out-of-domain data	62.5	51.8	54.5	49.5

Table 3. Results of distillation from intermediate-layer representations (%).

Elements	SST-2	MR	QNLI	RTE
MLM Logits	61.5	57.2	55.4	54.2
+ Top 4 layers	61.2	56.8	56.8	53.8
+ Skip layers	61.2	56.8	56.9	53.1

proaches for FT yield poor results, which are even worse than FT without KD (the lower bound). This is because the distilled models are severely overfitted to the training sets. ii) Prompt-KD achieves comparable performance to PT without KD, showing that the simple KD approach is not sufficient. iii) Boost-Distiller outperforms all the baselines by a large margin across all datasets. Other KD settings are shown in subsequent sections.

3.3. Detailed Analysis

We analyze Boost-Distiller in various aspects.

Ablation Study. Ablation results of Boost-Distiller are presented in Table 2. Due to the space limitation, we report the results over SST-2, MR, QNLI and RTE only. We can see that all the modules in Boost-Distiller contribute to the performance improvement. Yet, the degrees of improvement vary from tasks. For example, out-of-domain data is more important for SST-2, MR and QNLI, while teacher and fake logits play a vital role for RTE.

Results of Intermediate-layer KD. We further show that distilling intermediate layers is not always beneficial for few-shot KD. To achieve this, we present the KD results based on MLM Logits, as well as the hidden states from top 4 layers and skip layers, shown in Table 3. To make it fair for all the settings, we remove other parts of Boost-Distiller (such as fake logits) in all experiments. The results indicate that the KD performance drops over 3 out of 4 tasks, showing the difficulty of few-shot KD using intermediate representations. For simplicity, in Boost-Distiller, we do not employ any intermediate-layer KD loss.

Dataset Scale Analysis. We vary the number of training instances per class K from 16 to 512, and report the performance of Boost-Distiller and the method without KD (i.e., PET) in Figure 2. We can see that Boost-Distiller consistently outperforms PET with different Ks, showing that it can improve the model performance regardless of the choice of K. Hence, our work is beneficial for both few-shot and non-few-shot learning scenarios. Yet, it has a greater contribution when the training set is small.

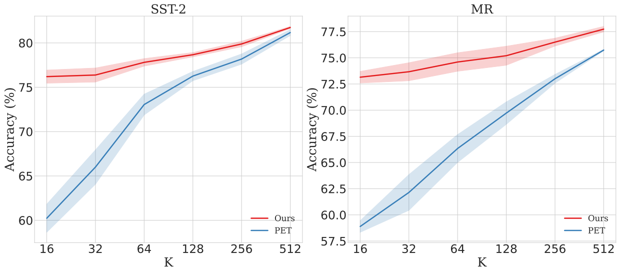


Fig. 2. Dataset scale analysis (%).

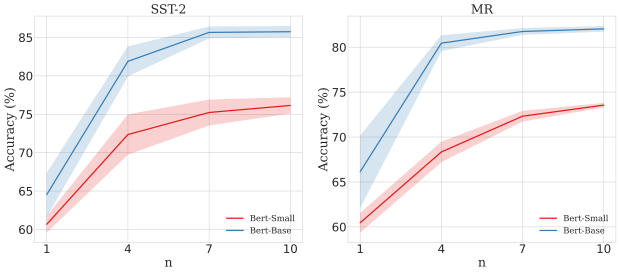


Fig. 3. Out-of-domain data analysis using BERT-base and BERT-small as student models (%).

Out-of-Domain Data and Model Analysis. We further vary the amount of out-of-domain training data ($n = 1, 4, 7, 10$) and employ BERT-base and BERT-small as student models. The results are shown in Figure 3. We find that the performance of both models improves with n becoming larger, showing that Boost-Distiller can be applied to various sizes of PLMs. The performance becomes relatively stable when $n \geq 7$.

4. CONCLUSION AND FUTURE WORK

In this work, we have presented Boost-Distiller, the first few-shot KD algorithm for prompt-based learners based on out-of-domain data. The proposed Boost-Distiller specifically considers heuristically-generated fake logits and cross-domain model logits weighted with domain expertise scores to improve the KD performance. Experimental results over a variety of NLP tasks and datasets show that the Boost-Distiller framework consistently outperforms strong baselines by a large margin.

Our work focuses on distilling models for Natural Language Understanding (NLU) tasks. It would also be possible to extend our work to Natural Language Generation (NLG) tasks, which will be addressed in future work.

Acknowledgment. This work is partially supported by the Alibaba Group through the Alibaba Innovative Research Program under Grant No. H20210412.

5. REFERENCES

[1] X. Han, Z. Zhang, N. Ding, Y. Gu, X. Liu, Y. Huo, J. Qiu, L. Zhang, W. Han, M. Huang, Q. Jin, Y. Lan, Y. Liu, Z. Liu, Z. Lu, X. Qiu, R. Song, J. Tang, J. Wen, J. Yuan, W. X. Zhao, and J. Zhu, "Pre-trained models: Past, present and future," CoRR, vol. abs/2106.07139, 2021.

[2] T. Shin, Y. Razeghi, R. L. L. IV, E. Wallace, and S. Singh, "Autoprompt: Eliciting knowledge from language models with automatically generated prompts," in EMNLP, 2020, pp. 4222–4235.

[3] T. Schick and H. Schütze, "Exploiting cloze-questions for few-shot text classification and natural language inference," in EACL, 2021, pp. 255–269.

[4] T. Gao, A. Fisch, and D. Chen, "Making pre-trained language models better few-shot learners," in ACL/IJCNLP, 2021, pp. 3816–3830.

[5] X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang, and J. Tang, "GPT understands, too," CoRR, vol. abs/2103.10385, 2021.

[6] C. Wang, J. Wang, M. Qiu, J. Huang, and M. Gao, "Transprompt: Towards an automatic transferable prompting framework for few-shot text classification," in EMNLP, 2021, pp. 2792–2802.

[7] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in NAACL-HLT, 2019, pp. 4171–4186.

[8] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," Int. J. Comput. Vis., vol. 129, no. 6, pp. 1789–1819, 2021.

[9] R. Tang, Y. Lu, L. Liu, L. Mou, O. Vechtomova, and J. Lin, "Distilling task-specific knowledge from BERT into simple neural networks," CoRR, vol. abs/1903.12136, 2019.

[10] S. Sun, Y. Cheng, Z. Gan, and J. Liu, "Patient knowledge distillation for BERT model compression," in EMNLP-IJCNLP, 2019, pp. 4322–4331.

[11] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu, "Tinybert: Distilling BERT for natural language understanding," in EMNLP (Findings), 2020, pp. 4163–4174.

[12] S. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, and H. Ghasemzadeh, "Improved knowledge distillation via teacher assistant," in AAAI, 2020, pp. 5191–5198.

[13] S. Sun, Z. Gan, Y. Fang, Y. Cheng, S. Wang, and J. Liu, "Contrastive distillation on intermediate representations for language model compression," in EMNLP, 2020, pp. 498–508.

[14] L. Yuan, F. E. H. Tay, G. Li, T. Wang, and J. Feng, "Revisiting knowledge distillation via label smoothing regularization," in CVPR, 2020, pp. 3902–3910.

[15] A. Williams, N. Nangia, and S. R. Bowman, "A broad-coverage challenge corpus for sentence understanding through inference," in NAACL-HLT, 2018, pp. 1112–1122.

[16] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," in EMNLP, 2015, pp. 632–642.

[17] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100, 000+ questions for machine comprehension of text," in EMNLP, 2016, pp. 2383–2392.

[18] M. O. Dzikovska, R. D. Nielsen, C. Brew, C. Leacock, D. Giampiccolo, L. Bentivogli, P. Clark, I. Dagan, and H. T. Dang, "Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge," in SemEval@NAACL-HLT, 2013, pp. 263–274.

[19] W. B. Dolan and C. Brockett, "Automatically constructing a corpus of sentential paraphrases," in IWP@IJCNLP, 2005.

[20] B. Pang and L. Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," in ACL, K. Knight, H. T. Ng, and K. Oflazer, Eds., 2005, pp. 115–124.

[21] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in EMNLP, 2013, pp. 1631–1642.

[22] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," CoRR, vol. abs/1907.11692, 2019.

[23] I. Turc, M. Chang, K. Lee, and K. Toutanova, "Well-read students learn better: The impact of student initialization on knowledge distillation," CoRR, vol. abs/1908.08962, 2019.