Intra-Class Similarity-Guided Feature Distillation

Khouloud Saadi[1] **Jelena Mitrović** [1,2] **Michael Granitzer**[1]
[1]University of Passau, Germany
[2]Institute for Artificial Intelligence Research and Development of Serbia, Serbia
{Khouloud.Saadi, Jelena.Mitrovic, Michael.Granitzer}@uni-passau.de

Abstract

Knowledge Distillation (KD) is an effective technique for compressing large language models through the teacher-student framework. Previous work in feature distillation mainly applied an exact matching between the hidden representations of the student and the teacher. However, as the student has a lower capacity compared to the teacher, it may struggle to mimic its exact hidden representations. This leads to a large discrepancy between their features as shown in preceding research. Therefore, we propose intra-class similarity-guided feature distillation, a novel approach to make the task easier for the student. In this work, we map each sample representation by the student to its K nearest neighbor samples representations by the teacher that are within the same class. This method is novel and can be combined with other distillation techniques. Empirical results show the effectiveness of our proposed approach by maintaining good performance on benchmark datasets.

1 Introduction

Knowledge distillation (KD) [Romero et al., 2014, Hinton et al., 2015] is known as an effective technique to compress large language models (LLMs) [Sun et al., 2019, Sanh et al., 2019, Jiao et al., 2020]. It is a framework to train a student network, the model with fewer parameters, to mimic the behavior of a teacher network, the over-parameterized model, on a group of data points. There are different approaches of knowledge distillation where the teacher is dynamic as in [Zhou et al., 2021, Ma et al., 2022] or static as in [Jiao et al., 2020, Sun et al., 2019]. The knowledge embedded in various components of the teacher can be distilled to the student. As examples, we can mention the prediction layer [Sanh et al., 2019, Hinton et al., 2015], the attention matrices [Jiao et al., 2020, Wang et al., 2021], and the hidden states [Sun et al., 2019, Saadi et al., 2023, Jiao et al., 2020]. In [Kovaleva et al., 2019], it is shown that LLMs, e.g., BERT, suffer from over-parametrization in domain-specific tasks. Thus, task-specific distillation has been an active research topic. In this work, we mainly focus on task-specific feature distillation from a static teacher.

Existing methods in feature distillation tried to improve the loss function where MSE [Sun et al., 2019, Jiao et al., 2020], cosine distance [Sanh et al., 2019], and correlation function [Saadi et al., 2023] are used to match the hidden representations of the teacher and the student. However, previous work mostly applied a one-to-one mapping between the student hidden representations and the teacher hidden representations [Sun et al., 2019, Sanh et al., 2019] neglecting the capacity gap between them. In fact, each sample representation by the student is mapped to the same exact sample representation by the teacher. Nevertheless, as detailed in [Chen et al., 2022], in layer distillation, the student may struggle to mimic the hidden representations of the teacher because of their large capacity difference. This always results in huge discrepancies between their feature representations. Furthermore, as shown in [Liang et al., 2023], training a student to achieve discriminative feature extraction for the main classification task and exact feature matching for distillation at the same time, is considered a multi-task learning. It is also shown that, in this case, it tends to over-fit the teacher's hidden states representations.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

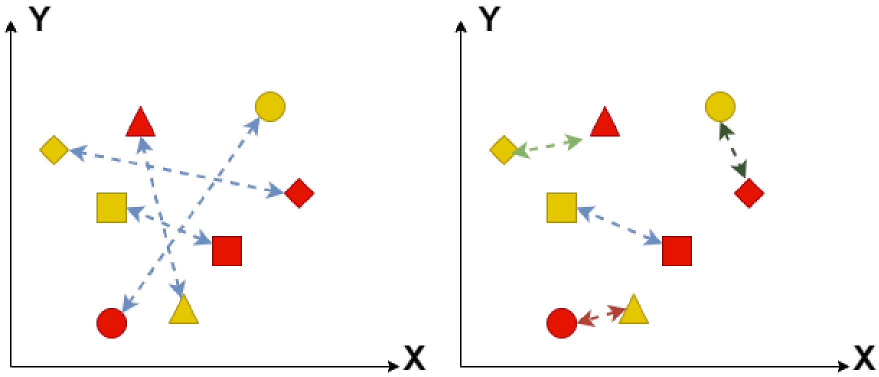

Figure 1: Left: Typical feature distillation. Right: Our proposed approach. For simplicity, we set K = 1. The arrows represents the loss per sample. Red shapes represent the teacher samples representations. Yellow shapes represent the student samples representations. The same samples are marked with the same shapes. The samples in the figure are from the same class

Motivated by this, we propose intra-class similarity-guided feature distillation, a novel approach where we introduce a new mapping between the student and teacher hidden representations. In fact, we match each student's sample representation with its K nearest neighbor teacher's samples representations which are within the same class. This new mapping will reduce the difficulty of the distillation task for the student model. Furthermore, we can look at our new mapping as a relaxation for the feature distillation task, so the student will not overfit the teacher features as detailed in [Liang et al., 2023]. Instead, it will focus better on the main feature extraction task while utilizing the teacher features as guidance.

In Figure 1, we illustrate the key idea of our approach using a simple example. In the left side, we present the typical features matching approach where each student sample representation is mapped to its exact sample representation by the teacher . In the right side, we present our new proposed approach where the mapping is done between each student sample representation and its nearest sample representation, from the same class, by the teacher. In the existing approach (Right), as sometimes the student's sample representation is very far from the teacher same sample representation, it is hard for the student to match it with its lower capacity, unlike in our proposed approach (Left) where we try to minimize the shortest distances taking advantages of the intra-class similarities.

In this work, we distill the last hidden representation of the teacher to the student as in [Tian et al., 2019, Yang et al., 2020] where we try to group together the samples representations of the same class, revealing the intra-class similarities. Mainly, because it is the closest to the classifier and will immediately affect the classification performance [Yang et al., 2020]. We also assume that the teacher's last hidden state and the student's last hidden state have the same dimension.

2 Methodology

Different from previous feature distillation work which applies a sample-wise representation alignment, we propose a KNN-based feature KD, a novel feature distillation method where the alignment is done between each sample representation by the student and its K nearest neighbors representations by the teacher which are from the same class . Our approach makes the task easier for the student. Moreover, As illustrated in Figure 2, the average intra-class similarity across the 4 GLUE benchmark datasets is higher with our method compared to the typical layer distillation technique. This highlighting the effect of our approach in learning more compact class-embedding. To empirically verify this hypothesis, we compute the intra-class cosine similarity M_{ICS} as following:

$$M_{ICS} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{c_i} \frac{< s_i \cdot s_j >}{c_i ||s_i||_2 ||s_j||_2}$$

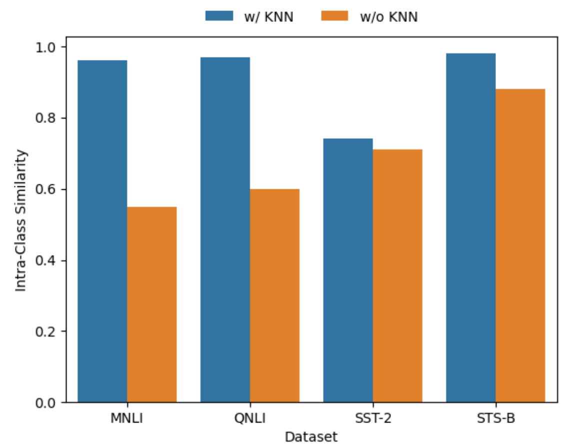

Figure 2: Intra-Class Similarity: Our approach VS typical feature distillation

N is the batch-size, s_j is the j-th sample belonging to the same class of s_i, and c_i is the total number of s_j in the batch of size N.

Typically, in a KD framework, we have the over-parameterized knowledgeable teacher modeled by f_θ. The efficient student network is modeled by $g_{\theta'}$ which has a lower number of parameters compared to the teacher $|\theta^{'}| << |\theta|$. An input batch X is fed to f_θ and $g_{\theta'}$ simultaneously to produce the last hidden representations Y_t and Y_s, respectively. Usually, to perform the feature distillation task, an MSE is computed between Y_t and Y_s [Sun et al., 2019, Jiao et al., 2020]. In fact, each sample representation in Y_s is mapped to its representation in Y_t. In this work, we propose a novel mapping approach to reduce the difficulty of the task for the student. We propose to map each sample representation in Y_s to its K nearest neighbors, that have the same label, in Y_t. In details, given a sample x in the input batch X where the batch X contains N samples. Its student representation r^S is with dimension n. $r^S = g_{\theta'}(x)$. $F = \{s \mid s \in X, \text{and label}(x) = \text{label}(s)\}$ contains the elements in the batch with the same label as x. $G = \{d \mid d = \sum_{j=1}^{n} \left(f_\theta(s)_j - r_j^S\right)^2, \text{and } s \in F\}$ contains the distances between each sample s in F and x. $G_K = \{i_1, i_2, i_3, ..., i_K \mid d_{i_1} < d_{i_2} < d_{i_3} ... < d_{i_K}, \text{and } K < N\}$ contains the indices of the K nearest points to x. The feature KD loss per sample is:

$$l_{hidd}(x) = \frac{1}{n} \sum_{k \in G_K} \sum_{j=1}^{n} \left(f_\theta(s_k)_j - g_{\theta'}(x)_j\right)^2$$

The final feature KD loss over all the batch samples is computed as following:

$$L_{hidd} = \sum_{x \in X} l_{hidd}(x)$$

The final KD loss is computed as following:

$$L_{KD} = \alpha_1 L_{hidd} + \alpha_2 L_{soft}$$

The final training loss of the student is computed as following:

$$L = L_{KD} + \alpha_3 L_{CE}$$

α_1, α_2, and α_3 are the contributions of the 3 loss components to the final training loss. L_{soft} is the logit distillation loss as in [Sanh et al., 2019, Jiao et al., 2020], which is the temperated KL divergence between the student logits and the teacher logits. L_{CE} is the cross entropy loss between the ground truth labels and the student predictions.

3 Experimental Results

3.1 Experimental Setup

Datasets In this work, we evaluate our proposed method on the validation set of 7 GLUE benchmark datasets [Wang et al., 2018]. The GLUE dataset is the typical benchmark for Knowledge distillation in NLP [Zhou et al., 2021]. It is composed of several datasets for different tasks. In our evaluation, we use MNLI, QNLI, and RTE for natural language inference; SST-2 is used for sentiment classification; QQP, MRPC, and STS-B are used for paraphrase similarity matching. The reported results are in the same format as on the official GLUE leader board.

Baselines In this work, the teacher is a 12-layer BERT-base-uncased model, fine-tuned on each GLUE task, with 110M parameters distilled into a 6-layer BERT_6 student model with 66M parameters. The number of epochs, the sequence length, the batch size, the learning rate are set to 5, 128, 32, and $\{1e-5, 3e-5, 5e-5\}$, respectively for the teacher fine tuning. We compare our proposed method with different state-of-the-art BERT compression approaches, including DistilBERT [Sanh et al., 2019], BERT-PKD [Sun et al., 2019], PD[Turc et al., 2019], TinyBERT [Jiao et al., 2020], BERT-of-Theseus [Xu et al., 2020], MetaDistil [Zhou et al., 2021], MiniLM v2 [Wang et al., 2021], and ReptileDistil [Ma et al., 2022]

Training settings For the baseline methods we report the same results in [Ma et al., 2022], which are from the corresponding original paper. In our work, following [Ma et al., 2022, Jiao et al., 2020], we initialize the student with the general TinyBERT_6 model weights. Similar to [Ma et al., 2022], the sequence length, the batch size, the number of epochs, and the temperature are set to 128, 32, 5, and 5, respectively. Similar to [Sanh et al., 2019, Jiao et al., 2020], α_2 and α_3 are set to 0.5 and 0.5, respectively. Following [Sun et al., 2019, Zhou et al., 2021, Ma et al., 2022], we conduct a grid search over student learning rate from $\{1e-5, 3e-5, 5e-5\}$, the K (number of nearest neighbors) from $\{1, 2, 3, 5\}$, and α_1 from $\{0.1, 0.01, 0.001\}$ and save the best model. All the experiments are repeated for 4 random seeds as in [Sun et al., 2019] and the average is reported.

3.2 Results

In this section, we discuss the experimental results of our approach.

Method	SST-2 (67k) Acc	MRPC (3.7k) F1/Acc	STS-B (5.7k) Pear/Spea	QQP (364k) F1/Acc	MNLI (393k) Acc m/mm	QNLI (105k) Acc	RTE (2.5k) Acc
BERT_{BASE} [Devlin et al., 2019]	93.0	91.6/87.6	90.2/89.8	88.5/91.4	84.6/84.9	91.2	71.4
DistilBERT [Sanh et al., 2019]	91.3	87.5/-	-/86.9	-/88.5	82.2/-	89.2	59.9
BERT-PKD [Sun et al., 2019]	91.3	85.7/-	-/86.2	-/88.4	81.3/-	88.4	66.5
PD [Turc et al., 2019]	91.1	89.4/84.9	-	87.4/90.7	82.5/83.4	89.4	66.7
TinyBERT [Jiao et al., 2020]	**93.0**	90.6/86.3	**90.1/89.6**	88.0/91.1	84.5/84.5	**91.1**	73.4
BERT-of-Theseus [Xu et al., 2020]	91.5	89.0/-	-/88.7	-/89.6	82.3/-	89.5	68.2
MiniLM v2 [Wang et al., 2021]	92.4	88.9/-	-	-/91.1	84.2/-	90.8	69.4
MetaDistil [Zhou et al., 2021]	92.3	91.1/86.8	89.4/89.1	**88.1/91.0**	83.5/83.8	90.4	72.1
ReptileDistil [Ma et al., 2022]	92.2	91.6/87.7	89.5/89.3	87.6/90.1	83.7/83.7	90.5	75.3
Ours	92.5	**92.5/89.64**	89.7/89.5	87.7/90.9	**84.5/84.5**	90.8	**75.8**

Table 1: Experimental results on the development set of GLUE. The numbers and the strings under each dataset name indicated the number of samples and the metrics.

As shown in Table 1, our proposed approach outperforms all the state-of-the-art methods on three datasets i.e., MRPC, MNLI, and RTE. While we distill the knowledge from a static teacher, ours outperforms both KD state-of-the-art MetaDistil and ReptileDistil, where the teacher is dynamic, on most of the datasets. While we distill the knowledge only from the last hidden representation of the teacher, ours outperforms BERT-PKD on all the datasets, which distills several hidden representations from the teacher to the student. It is also worth mentioning that, although in [Wang et al., 2023], the authors showed that the attention distillation is the best performing objective, ours outperforms MiniLM v2, which distills the attention, on all the datasets and TinyBERT, which distills the attention, all the hidden states, and the logits, on 3 datastets.

4 Conclusion

In this paper, we introduced a new mapping between the hidden representations of the teacher and the student. In fact, each sample representation by the student is mapped to its K nearest neighbors representations by the teacher. Our approach makes the task easier for the student and helps it to learn more compact samples representations. Empirical results showed the effectiveness of our proposed method. Future work will include exploring adding a projector to dispose of the requirement that the student and the teacher must have the same last hidden states dimension.

Acknowledgment

The project on which this report is based was funded by the German Federal Ministry of Education and Research (BMBF) under the funding code 01IS20049. The author is responsible for the content of this publication.

SPONSORED BY THE

Federal Ministry
of Education
and Research

References

Yudong Chen, Sen Wang, Jiajun Liu, Xuwei Xu, Frank de Hoog, and Zi Huang. Improved feature distillation via projector ensemble. *Advances in Neural Information Processing Systems*, 35: 12084–12095, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423`.

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

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.372. URL `https://aclanthology.org/2020.findings-emnlp.372`.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. Revealing the dark secrets of bert. *arXiv preprint arXiv:1908.08593*, 2019.

Chen Liang, Simiao Zuo, Qingru Zhang, Pengcheng He, Weizhu Chen, and Tuo Zhao. Less is more: Task-aware layer-wise distillation for language model compression. In *International Conference on Machine Learning*, pages 20852–20867. PMLR, 2023.

Xinge Ma, Jin Wang, Liang-Chih Yu, and Xuejie Zhang. Knowledge distillation with reptile meta-learning for pretrained language model compression. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4907–4917, 2022.

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

Khouloud Saadi, Jelena Mitrović, and Michael Granitzer. Learn from one specialized sub-teacher: One-to-one mapping for feature-based knowledge distillation. In *ICML 2023 Workshop Neural Compression: From Information Theory to Applications*, 2023.

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

S. Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. Patient knowledge distillation for bert model compression. In *Conference on Empirical Methods in Natural Language Processing*, 2019. URL `https://api.semanticscholar.org/CorpusID:201670719`.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*, 2019.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL `https://aclanthology.org/W18-5446`.

Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. MiniLMv2: Multi-head self-attention relation distillation for compressing pretrained transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2140–2151, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.188. URL `https://aclanthology.org/2021.findings-acl.188`.

Xinpeng Wang, Leonie Weissweiler, Hinrich Schütze, and Barbara Plank. How to distill your bert: An empirical study on the impact of weight initialisation and distillation objectives. *arXiv preprint arXiv:2305.15032*, 2023.

Canwen Xu, Wangchunshu Zhou, Tao Ge, Furu Wei, and Ming Zhou. BERT-of-theseus: Compressing BERT by progressive module replacing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7859–7869, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.633. URL `https://aclanthology.org/2020.emnlp-main.633`.

Jing Yang, Brais Martinez, Adrian Bulat, and Georgios Tzimiropoulos. Knowledge distillation via softmax regression representation learning. In *International Conference on Learning Representations*, 2020.

Wangchunshu Zhou, Canwen Xu, and Julian McAuley. Bert learns to teach: Knowledge distillation with meta learning. In *Annual Meeting of the Association for Computational Linguistics*, 2021. URL `https://api.semanticscholar.org/CorpusID:237250417`.