Learn From One Specialized Sub-Teacher: One-to-One Mapping for Feature-Based Knowledge Distillation

Khouloud Saadi [1] Jelena Mitrović [1] Michael Granitzer [1]

Abstract

Knowledge Distillation is known as an effective technique to compress over-parameterized language models. In this work, we propose to break down the global feature distillation task into N local sub-tasks. In this new framework, we consider each neuron in the last hidden layer of the teacher network as a specialized sub-teacher. We also consider each neuron in the last hidden layer of the student network as a focused sub-student. We make each focused sub-student learn from one corresponding specialized sub-teacher and ignore the others. This will facilitate the task for the sub-student and keep him focused. This method is novel and can be combined with other distillation techniques. Empirical results show that our proposed approach outperforms the state-of-the-art methods by maintaining higher performance on most benchmark datasets.

1. Introduction

Large language models, also known as general purpose language models, have revolutionized the NLP domain (Devlin et al., 2018; Brown et al., 2020). They are large architectures composed of several transformer blocks (Vaswani et al., 2017), typically trained on large unlabeled corpora in a self-supervised way (Devlin et al., 2018; Brown et al., 2020). They achieved state-of-the-art (SOTA) performance on downstream tasks through fine-tuning when the data is scarce (Devlin et al., 2018). However, as these models are usually large, e.g., BERT has millions of parameters (Devlin et al., 2018) and GPT-3 has billions of parameters (Brown et al., 2020), they are not highly adapted to real-world applications. Model compression is an active area of research, where the model size is effectively reduced without a significant loss of performance (Xu & McAuley, 2022).

Knowledge Distillation (KD) by (Hinton et al., 2015) is one of the effective compression techniques in NLP where the knowledge of a highly capable large model, i.e., teacher, is transferred to a smaller model, i.e., student. KD essentially requires designing a loss function to minimize the distance of the output or the intermediate representations between the student and the teacher (Sanh et al., 2019). To distill the intermediate representations, previous research relied on the mean square error (MSE) as an objective function between the student and the teacher global representations (Sun et al., 2019; Jiao et al., 2019). However, this metric is sensitive to scale (Saadi & Taimoor Khan, 2022) and it is not accurate in high dimensional space(Aggarwal et al., 2001; Houle et al., 2010). Other works used the cosine distance as an alternative, but, it also has several limitations (Zhou et al., 2022; Schütze et al., 2008) such as not performing well with sparse data.

In this work, we propose a novel KD approach where we reformulate the feature distillation task from a global problem to N local sub-problems. Each unit, i.e., neuron, in the teacher's last hidden layer is in charge of distilling its knowledge to the corresponding unit in the student model. We call this a one-to-one matching between the two networks. In this new framework, we consider each neuron in the teacher's last hidden layer as a specialized sub-teacher and each neuron in the student's last hidden layer as a focused sub-student. To distill the knowledge from each specialized sub-teacher to each focused sub-student, we propose a local correlation-based objective function. Local, i.e., per-neuron basis. Empirical results show that studying the global feature distillation task from a local viewpoint helped the student to meet the global teacher's features representation. To sum up, our contributions are the following:

- We reformulate the global feature distillation task into N local sub-tasks where we do a one-to-one mapping between the model units.
- We propose a local correlation-based objective function for the distillation task.
- We conduct experiments on 8 GLUE datasets (Wang

[1] University of Passau, Germany. Correspondence to: Khouloud Saadi <Khouloud.Saadi@uni-passau.de>, Jelena Mitrović <Jelena.Mitrovic@uni-passau.de>, Michael Granitzer <Michael.Granitzer@uni-passau.de>.

Published at the ICML 2023 Workshop on Neural Compression: From Information Theory to Applications. Copyright 2023 by the author(s).

et al., 2018), the SQUAD V1, and the IMDB dataset, where our approach performs the best in most cases.

2. Related Work

Knowledge distillation has been proven as an effective technique for model compression. It can be applied during the pre-training stage to generate general-purpose distilled models (Jiao et al., 2019; Sanh et al., 2019) and during the fine-tuning stage to generate task-specific distilled models (Zhou et al., 2021; Liang et al., 2020).

In (Kovaleva et al., 2019), the authors prove that large language models, e.g., BERT, suffer from over-parametrization in domain-specific tasks. Thus, previous work has been improving the task-specific distillation. Several methods focused on enhancing the objectives of the distillation process. These improvements mainly focused on which part of the teacher architecture can be distilled into the student architecture such as the attention matrices (Jiao et al., 2019), the different hidden states (Sun et al., 2019), and the prediction layer (Hinton et al., 2015).

Coming up with effective metrics to distill the knowledge from any part of the teacher into the student is critical. For the logit-based KD, the KL divergence or the MSE are used as objective functions to minimize the distance between the logits of the student and the logits of the teacher (Hinton et al., 2015; Zhou et al., 2021). Following that, in (Zhao et al., 2022), the authors provided a novel viewpoint to study the logit distillation by reformulating the classical KL divergence loss into two parts, which showed a good improvement. In the feature-based KD, the MSE and the cosine distance are mainly used as objective functions to align the features representation of the teacher and the student (Sun et al., 2019; Jiao et al., 2019; Sanh et al., 2019).

In this work, we argue that relying on the MSE and the cosine distance as objective functions to align the intermediate representations between the student and the teacher is not an optimal choice. MSE is sensitive to scale and does not perform well in high-dimensional space. In a high dimensional space, which is the case in a neural network, the data tend to be sparse and all the data points become uniformly distant from each other (Aggarwal et al., 2001; Houle et al., 2010). In (Zhou et al., 2022), it also shows that the cosine distance is not an accurate measurement of similarity between BERT embeddings. Moreover, the cosine distance measure can be also affected by sparsity. In fact, in high-dimensional space, it can output large angles between two sparse vectors although they are similar in the non-zero components (Schütze et al., 2008). One other important limitation of the cosine distance is that it is a global metric. For example, let W_1 be the output tensor of a given layer in the teacher and W_2 be the output of the corresponding layer in the student. If $W_1 = [1, 5]$ and $W_2 = [5, 1]$, then the $CosD(W_1, W_2) = 1 - \dfrac{W_1 \cdot W_1}{\|W_1\|_2 \|W_1\|_2} = 0.6154$. It is high but in actual the layers learned the same representation.

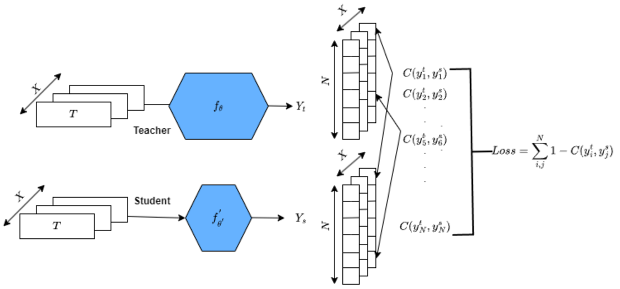

Figure 1. The scheme of our idea: X is the input batch. T is a given sample. C is the cross correlation function. The **Student** and the **Teacher** are modeled by $f'_{\theta'}$ and f_θ, respectively. Y_s and Y_t are the features representation of the last hidden layer of the student and the teacher, respectively.

Instead of treating the feature distillation as a global task, we reformulate it as multiple local sub-tasks. Furthermore, for each local KD sub-task, we propose a neuron-based correlation objective function that operates across batches. This improves the student's ability to meet the global representation provided by the teacher.

3. Methodology

In this work, we provide a new viewpoint on how to study the feature distillation task. We break down the global KD task into multiple local KD sub-tasks. Specifically, in the last hidden layer of the teacher, we consider each unit as a specialized sub-teacher. In the last hidden layer of the student, we also treat each unit as a focused sub-student. Each sub-student should concentrate and learn only from the corresponding specialized sub-teacher. We call this a one-to-one matching, which is reformulated as the cross-correlation between the outputs of each sub-teacher and each sub-student. Although our approach can be applied to different student-teacher hidden layers, as in (Sun et al., 2019), this work specifically focuses on the last layer.

Typically, in a KD framework, as illustrated in Figure 1, we have the teacher network modeled by f_θ, which is an over-parameterized knowledgeable model. The student network is modeled by $f'_{\theta'}$ which has a lower number of parameters compared to the teacher. The input batch X is fed to f_θ and $f'_{\theta'}$ simultaneously to produce the batches of features representation Y_t and Y_s, respectively.

Assuming that Y_t and Y_s are the features representation of the last hidden layer of the teacher h_t^{Last} and the last hidden layer of the student h_s^{Last}, respectively. Y_s and Y_t are assumed to be mean-centered over the batch dimension. Assuming that h_t^{Last} and h_s^{Last} have N hidden units. The unit_i in h_t^{Last} represents the sub-$\mathrm{teacher}_i$ and the unit_i in h_s^{Last} represents the sub-$\mathrm{student}_i$. The sub-task_i is distilling the Knowledge from the sub-$\mathrm{teacher}_i$ to the sub-$\mathrm{student}_i$ by reducing the distance between the features learned by each of them. To simplify the task for the sub-$\mathrm{student}_i$ and to keep him focused, we force him to learn only from the sub-$\mathrm{teacher}_i$ and ignore the other teachers, i.e., neurons. We reformulate the objective function in this one-to-one mapping as maximizing the cross-correlation between the two variables y_i^t and y_i^s. y_i^t and y_i^s are the output values of the sub-$\mathrm{teacher}_i$ and the sub-$\mathrm{student}_i$, respectively. The variables y_i^t and y_i^s have X samples coming from the different examples in the input batch. Maximizing the correlation across batches between the two aforementioned variables, i.e., minimizing the following loss function:

$$L_i = (1 - C_{ii})^2$$

where C_{ii} is the cross-correlation value between the variables y_i^t and y_i^s:

$$C_{ii} = \frac{\sum_{b=1}^{|X|} y_{b,i}^t y_{b,i}^s}{\sqrt{\sum_{b=1}^{|X|} (y_{b,i}^t)^2}\sqrt{\sum_{b=1}^{|X|} (y_{b,i}^s)^2}}$$

b is the index of a given sample in the input batch X. i, j index the output dimension of the last hidden layer in both the teacher and the student. In fact, i is the index of the i_{th} element in the output and it is also the index of the i_{th} sub-teacher or sub-student in the last hidden layer. As we want to make the task easier for the sub-$\mathrm{student}_i$, so he can effectively digest the received information, we force him to mimic only the $\mathrm{teacher}_i$ and ignore all the rest. This results in minimizing the following term:

$$R_i = \sum_{j \neq i}^{N} C_{ij}^2$$

Thus, our final local KD loss function is:

$$L_i^F = (1 - C_{ii})^2 + \sum_{j \neq i}^{N} C_{ij}^2$$

In L_i^F, the first term, is for maximizing the cross-correlation over batches between the output of the sub-$\mathrm{student}_i$ and the output of the sub-$\mathrm{teacher}_i$. The second term is for minimizing the cross-correlation between the sub-$\mathrm{student}_i$ and each sub-$\mathrm{teacher}_j$ given $j \neq i$ and $j \in \{1, 2, ..., N\}$. Our end distillation loss is the sum of the N local KD losses.

$$L_{KD} = \sum_{i}^{N} L_i = \sum_{i}^{N} (1 - C_{ii})^2 + \sum_{i}^{N} \sum_{j \neq i}^{N} C_{ij}^2$$

Additionally, we introduce λ_1 and λ_2 as the weights to control the contribution of the first term and the second term of the loss function, respectively:

$$L_{KD} = \lambda_1 \sum_{i}^{N} (1 - C_{ii})^2 + \lambda_2 \sum_{i}^{N} \sum_{j \neq i}^{N} C_{ij}^2$$

The final training loss of the original student is:

$$L = \alpha L_{CE} + \beta L_{KD}$$

Where L_{CE} is the classical cross entropy loss between the student predictions and the ground truth labels. Empirically, our one-to-one mapping effectively facilitated the alignment of the student representation with the global teacher representation. Our approach can be implemented with low cost and can be combined with other KD methods.

4. Experimental Results

In our KD framework, the teacher is the BERT-base model, with 110 million parameters, after being fine-tuned on each of the datasets for 3 epochs. The student is the DistilBERT-base with 66 million parameters. N is equal to 768. All experiments are repeated for 5 random seeds, the maximum sequence length is set to 128, and the batch size is set to 16.

4.1. Stand-Alone Experiments

In this stand-alone performance evaluation, we experiment with the SQUAD-V1 and the IMDB datasets. We distill the last hidden layer representation of the teacher to the student. We add our designed KD loss, the MSE as in (Sun et al., 2019), and the cosine distance as in (Sanh et al., 2019), as stand-alone regularizers to the hard loss between the student predictions and the ground truth labels. This will show the effectiveness of our proposed objective function for the feature distillation task. We experiment with 3 and 10 epochs. The weight of each KD stand-alone loss is fixed to 1 (Sanh et al., 2019; Jiao et al., 2019).

In Table 1, BERT_{12} and $\mathrm{Distilbert}_6$ refer to BERT-base with 12 transformer blocks and DistilBERT-base with 6 transformer blocks, respectively. $\mathrm{Distilbert}_6$-FT stands for fine-tuning the student without any distillation. $\mathrm{Distilbert}_6$-cosD stands for feature distillation with cosine distance objective. $\mathrm{Distilbert}_6$-MSE stands for feature distillation with MSE. We note that our proposed method achieves the best results on the feature distillation task. It could beat MSE and cosine distance, with a high range, on the squad and the imdb datasets. The results show that our approach performs best when the distillation task is run for 3 epochs. It has 78.79% and 86.95% as Exact Match (EM) and F1 on the squad dataset, respectively. It also has 93.87% as accuracy on the imdb dataset. For 10 epochs, while the performance

Table 1. Stand-Alone regularizers: SQUAD-V1: The evaluation reported as Exact Match (EM) and F1 on the dev set. IMDB: The evaluation reported as accuracy on the test set. Results are the average and the standard deviation of 5 random seeds. **Left**: Distillation was run for 3 epochs. **Right**: Distillation was run for 10 epochs.

Approach	SQUAD-V1 (%)	IMDB (%)
BERT$_{12}$(teacher)	80.36/88.13	94.06
Distilbert$_6$-FT	77.43±0.22/85.67±0.10	93.19±0.09
Distilbert$_6$-cosD	77.82±0.29/85.92±0.23	93.66±0.08
Distilbert$_6$-MSE	77.72±0.21/85.86±0.12	93.49±0.08
Distilbert$_6$-**ours**	**78.79±0.12/86.95±0.06**	**93.87±0.01**

Approach	SQUAD-V1 (%)	IMDB (%)
BERT$_{12}$(teacher)	80.36/88.13	94.06
Distilbert$_6$-FT	74.32±0.37/83.69±0.31	92.71±0.08
Distilbert$_6$-cosD	75.78±0.27/84.57±0.11	93.90±0.08
Distilbert$_6$-MSE	75.17±0.18/84.25±0.15	93.82±0.07
Distilbert$_6$-**ours**	**79.46±0.08/87.48±0.05**	**93.96±0.01**

Table 2. Results on the GLUE dataset: Evaluation reported on the dev set as the average of 5 random seeds. All values are in (%).

Approach	MRPC	RTE	CoLA	SST-2	STS-B	MNLI	QNLI	QQP
BERT$_{12}$(teacher)	87.86	66.79	54.84	90.02	89.30	82.85	90.70	88.04
Distilbert$_6$-FT	84.48	56.17	44.28	89.54	85.40	80.41	86.71	87.79
Distilbert$_6$-CosD	85.38	63.75	46.82	89.79	85.59	80.51	88.62	87.78
Distilbert$_6$-MSE	86.04	62.38	47.35	89.79	85.42	80.12	88.13	87.90
Distilbert$_6$-PKD	86.06	62.24	47.28	90.05	**85.77**	81.62	87.16	**88.33**
Distilbert$_6$-KD	84.56	56.53	45.86	89.63	85.50	80.51	87.71	87.60
Distilbert$_6$-**ours**	**86.57**	**63.83**	**50.73**	**90.44**	85.66	**82.76**	**89.54**	88.04

of the other approaches decreased, ours effectively increased. It achieves 79.46% and 87.48% as EM and F1, respectively, on the squad dataset and 93.96% on the imdb dataset. This proves that our approach facilitates the convergence of the student's representation to the teacher's representation and results in a better-generalized student. Another noteworthy aspect, is that the standard deviation of our approach is always low which indicates the stability and consistency of our resulting student model compared to others.

4.2. Comparison With Competing Methods

In this evaluation part, we compare the performance of our proposed approach with the competing methods on the commonly used GLUE benchmark dataset (Wang et al., 2018) for knowledge distillation in NLP. For comparison, we fine-tune the student on the tasks without distillation. We report the stand-alone results for MSE and cosine distance. We also generate the results of vanilla KD (Hinton et al., 2015) and PKD (Sun et al., 2019). For the PKD approach, we use a similar setting to (Sun et al., 2019). For the rest of the baselines, similar to (Sanh et al., 2019; Zhou et al., 2021), the number of epochs is set to 3. The temperature is set to 2 for the vanilla-KD (Hinton et al., 2015). All weights in loss functions are fixed to 1 except for the vanilla-KD, loss weight is chosen from $\{0.4, 0.5, 0.6\}$ (Sun et al., 2019) and 0.4 is the best match. In our approach, λ_1, λ_2, α, and β are set to 1, 5.10^{-3}, 0.5, and 5.10^{-3}, respectively. These values are found after running a search on the best hyper-parameter value. The number of epochs is set to 3.

The GLUE benchmark dataset is composed of several sub-datasets of different tasks. QQP, MRPC, and STS-B are for paraphrase similarity matching. SST-2 is for sentiment classification. MNLI, QNLI, and RTE are for natural language inference. CoLA is for linguistics acceptability. For MRPC and QQP we report the combined score from F1 and accuracy. For STS-B we report the combined score from Pearson and Spearman correlations. For CoLA we report the Matthew's correlation. For the rest of the tasks, accuracy is the metric.

As shown in Table 2, our approach outperforms KD, which stands for vanilla-KD (Hinton et al., 2015), and the PKD baselines on most of the GLUE tasks. Although our distillation objective is applied as a stand-alone between the last hidden layer of the student and the teacher, it could beat PKD. Note that PKD uses the MSE objective between several layers of the student and teacher networks. It also includes the vanilla-KD in its final loss. This reflects the effectiveness of our proposed approach.

5. Conclusion and Future Work

In this paper, we reformulated the global feature distillation problem into N local sub-problems. We proposed a one-to-one matching between each neuron in the last hidden layer of the teacher, i.e., specialized sub-teacher, and each neuron in the last hidden layer of the student. i.e., focused sub-student. To achieve this goal, we proposed a local correlation-based loss. Our approach only requires the teacher and the student to have the same last hidden layer size. Several experiments proved the effectiveness and consistency of our method. It is also worth mentioning that our approach can be added to any KD method in NLP or vision. Future work includes exploring the same distillation process with several intermediate layers.

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 01—S20049. The author is responsible for the content of this publication.

SPONSORED BY THE
Federal Ministry
of Education
and Research

References

Aggarwal, C. C., Hinneburg, A., and Keim, D. A. On the surprising behavior of distance metrics in high dimensional space. In *International conference on database theory*, pp. 420–434. Springer, 2001.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

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

Houle, M. E., Kriegel, H.-P., Kröger, P., Schubert, E., and Zimek, A. Can shared-neighbor distances defeat the curse of dimensionality? In *Scientific and Statistical Database Management: 22nd International Conference, SSDBM 2010, Heidelberg, Germany, June 30–July 2, 2010. Proceedings 22*, pp. 482–500. Springer, 2010.

Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., and Liu, Q. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*, 2019.

Kovaleva, O., Romanov, A., Rogers, A., and Rumshisky, A. Revealing the dark secrets of bert. *arXiv preprint arXiv:1908.08593*, 2019.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Liang, K. J., Hao, W., Shen, D., Zhou, Y., Chen, W., Chen, C., and Carin, L. Mixkd: Towards efficient distillation of large-scale language models. *arXiv preprint arXiv:2011.00593*, 2020.

Saadi, K. and Taimoor Khan, M. Effective prevention of semantic drift in continual deep learning. In *Intelligent Data Engineering and Automated Learning–IDEAL 2022: 23rd International Conference, IDEAL 2022, Manchester, UK, November 24–26, 2022, Proceedings*, pp. 456–464. Springer, 2022.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Schütze, H., Manning, C. D., and Raghavan, P. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge, 2008.

Sun, S., Cheng, Y., Gan, Z., and Liu, J. Patient knowledge distillation for bert model compression. *arXiv preprint arXiv:1908.09355*, 2019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

Xu, C. and McAuley, J. A survey on model compression for natural language processing. *arXiv preprint arXiv:2202.07105*, 2022.

Zhao, B., Cui, Q., Song, R., Qiu, Y., and Liang, J. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 11953–11962, 2022.

Zhou, K., Ethayarajh, K., Card, D., and Jurafsky, D. Problems with cosine as a measure of embedding similarity for high frequency words. *arXiv preprint arXiv:2205.05092*, 2022.

Zhou, W., Xu, C., and McAuley, J. Bert learns to teach: Knowledge distillation with meta learning. *arXiv preprint arXiv:2106.04570*, 2021.