The History and Geography Professor: A TinyBERT Application using Knowledge Distillation

Marco Colangelo[1*]

[1]Department of Computer Science, University of Milan, Milan, Italy.

Corresponding author(s). E-mail(s): marco.colangelo@studenti.unimi.it;

Abstract

This report contains the explanation of the work done for the final project of the Natural Language Processing course. The professor in charge of the course is Professor Alfio Ferrara, University of Milan. The project focused on the application of knowledge distillation, a technique used to compress large language models into smaller and more efficient versions. This is achived by training a smaller model (the student) to replicate the behavior of a larger model (the teacher). The main objective of this project was to implement and evaluate the TinyBERT model, a compact version of the BERT architecture, using knowledge distillation techniques. In particular, in this work, the BERT model (the teacher) acts as a geography and history professor that teaches to three versions of TinyBERT (the students), each of which has a different learning method from the teacher.

1 Introduction

Language model pre-training, such as BERT, has significantly improved the perfomances of many NLP tasks. However, pre-trained language models are usually computationally expensive, so it is difficult to efficiently execute them on resource-restricted devices. A novel Transformer distillation method, specially designed for knowledge distillation of the Transformer-based models, allows to effectively transfer to a smaller student model called TinyBERT the plenty of knowledge from a large pre-trained BERT model. A two-stage learning framework for TinyBERT which performs Transformer distillation at both the pretraining and task-specific learning stages

ensures that TinyBERT can capture the general-domain as well as the task-specific knowledge in BERT. [1]

1.1 State of the art work

So far, two main studies have defined the paradigm of model compression via knowledge transfer. The first, by Hinton et al. [2], introduced the seminal concept of Knowledge Distillation (KD), focusing on transferring the so-called dark knowledge from a large Teacher to a compact Student. Their approach minimizes the distance between the output probability distributions of the two networks, softened by a temperature parameter T. The second and more specialized study, by Jiao et al. [1], proposes a variant named TinyBERT, specifically designed for Transformer-based architectures. While Hinton's method operates on the final output layer, Jiao et al. argue that for complex language models, this is insufficient. Their method leverages a layer-to-layer distillation strategy, forcing the student to mimic not only the teacher's prediction but also its intermediate representations, such as attention matrices and hidden states.

2 Related Work

2.1 Aim of the work

This work has the aim to implement and evaluate the TinyBERT model as described in the second related work, trained with the presented two-stage learning framework to distill knowledge from a pre-trained BERT model. More specifically, the BERT model (the teacher) acts as a geography and history professor that teaches to three version of TinyBERT (the students), each of which has a different strategy to learn from the teacher.

- The first student is the "lazy" one: he simply learns to replicate everything the professor says, but does not replicate the reasoning used to formulate the answer. Specifically, he learns from the professor by minimizing the difference between his final probability distribution (the logits) and that of the professor. This is the type of learning most similar to that proposed by Hilton et al. [2]
- The second student is the "attentive" one: rather than focusing on the professor's final answer, he tries to understand what the professor pays attention to in a sentence, since attention captures linguistic and syntactic knowledge. Specifically, he learns from the professor by minimizing the MSE between his attention matrices and those of the professor.
- The third student is what can be called the "little professor": he learns to reason exactly like the professor, imitating his entire brain activity. Specifically, he learns from the professor by minimizing the error between his intermediate layer outputs and those of the professor. This is the approach proposed by Jiao et al. [1]

The three students are then evaluated by asking them a specific history or geography question and comparing their answers to the teacher's one, in order to understand which loss function allows the student to learn better from the teacher. We want to

verify if we can reach the results obtained in the lecterature [1], in which the third student outperforms the other two.

2.2 Data

Regarding the data used, the knowledge dataset used was the unlabeled corpus **WikiText** in its *WikiText-103* version, accessible through the **HuggingFace Datasets** library and loaded in streaming mode to reduce memory usage. WikiText was chosen because it is a collection of verified Wikipedia articles, with high syntactic and grammatical quality. All sequences are tokenized with `BertTokenizerFast` and truncated to a maximum length of 128 tokens. Unlike the original TinyBERT task-specific stage [1], our second stage does not rely on a supervised downstream dataset; it focuses on domain specialization under the masked language modeling objective.

2.2.1 Preprocessing

Since the professor teaches history and geography in this project, it was necessary to filter the dataset to retain only documents related to these subjects. A dictionary of representative terms for the two subjects was therefore defined, consisting of the following:

- History: "history", "empire", "war", "ancient", "king", "queen"
- Geography: "geography", "river", "mountain", "capital", "population"

During the streaming of the original dataset, each article was scanned. If at least one dictionary term was present, the document was retained in the training dataset; otherwise, it was discarded.

At this point, raw data was obtained and made digestible by the models. For the tokenization phase, we used pre-trained `BertTokenizerFast` and associated it with `bert-base-uncased`, because the student and teacher models must use the same numeric tokens to make a direct comparison of the outputs. All documents were truncated to a fixed length of 128 tokens to ensure uniform batches.

The final step of the preprocessing phase was to prepare the self-supervised learning using a **Masked Language Modeling (MLM)** technique using the `DataCollatorForLanguageModeling` class. We created a collator that randomly masks 15% of the tokens in each phase, replacing them with the special token `[MASK]`. The labels for the loss calculation are set such that the model only predicts the masked tokens.

2.3 Implementation

2.3.1 Structure of the models

For the teacher model, we used bert-base-uncased, with $N = 12, d = 768, h = 12$, and $109M$ parameters, where N is the number of layers, d is the length of the numerical vector that represents every single token inside the model and h is the number of attention heads. To optimize memory usage, the teacher's weights were frozen during training.

The TinyBERT model that implements the students, however, was built as in the literature with $M = 4, d' = 312, h = 12$, and $14.5M$ parameters, 78% fewer than the teacher [1]. To enable Hidden State Distillation, we added a linear projection layer W_h that maps the student's (312-dimensional) vector space to the teacher's (768-dimensional) vector space. We adopted a uniform mapping strategy ($g(m) = 3 \times m$) as defined in the lecterature [1], meaning that the 4 layers of the student learning from levels $3, 6, 9$ and 12 of the teacher.

2.3.2 Training phase

We adopt a TinyBERT-inspired two-stage distillation framework. More specifically:

1. In stage 1 we perform General Distillation to transfer broad linguistic knowledge from the teacher to the student using a general-purpose unlabeled corpus under the MLM objective.
2. In stage 2 we perform Domain-Adaptive Distillation, which continues training on a history/geography filtered subset of the corpus to encourage domain specialization.

This differs from the lecterature TinyBERT two-stage training [3], where stage 2 is a task-specific supervised distillation performed using a teacher fine-tuned on a downstream task (e.g., GLUE or SQuAD). In our setting, both stages remain MLM-based, and the specialization objective is achieved through domain filtering rather than supervised task labels.

For the experiments, a laptop equipped with a Nvidia GTX 1650 Max-Q GPU and 16GB of RAM was used to perform simple tests, while the main training phase involving the TinyBERT model was performed on a desktop PC equipped with a Nvidia RTX 3070 GPU and 32GB of RAM.

2.3.3 Distillation Objectives

All students are trained with a shared hard MLM loss to ensure they remain evaluable as MLMs. We combine this task with one or more distillation losses depending on the student variant:

$$\mathcal{L} = \lambda_{\text{hard}}\mathcal{L}_{\text{MLM}} + \lambda_{\text{pred}}\mathcal{L}_{\text{KD}} + \lambda_{\text{attn}}\mathcal{L}_{\text{attn}} + \lambda_{\text{hidn}}\mathcal{L}_{\text{hidn}} + \lambda_{\text{emb}}\mathcal{L}_{\text{emb}} \tag{1}$$

where:

- Hard MLM loss $\mathcal{L}_{\text{MLM}}$ is the cross-entropy on masked tokens only.
- Prediction-layer distillation $\mathcal{L}_{\text{KD}}$ is the knowledge divergence between teacher and student output distributions over the vocabulary, computed only on masked tokens using a temperature T.
- Attention distillation $\mathcal{L}_{\text{attn}}$ is the MSE between teacher and student pre-softmax attention scores.
- Hidden state distillation $\mathcal{L}_{\text{hidn}}$ is the MSE between aligned intermediate hidden representations.
- Embedding distillation $\mathcal{L}_{\text{emb}}$ is the MSE between aligned embedding outputs.

We train the three students sharing the same TinyBERT architecture and the same two-stage protocol, differentiating them only in the distillation signals they match.

During stage 1 (General Distillation)

- Student A (the lazy one): hard MLM only, so $\lambda_{\text{pred}} = 0, \lambda_{\text{attn}} = 0, \lambda_{\text{hidn}} = 0$
- Student B (the attentive one): hard MLM + attention-only distillation, so $\lambda_{\text{attn}} > 0$.
- Student C (the little professor): hard MLM + full intermediate distillation, so $\lambda_{\text{pred}}, \lambda_{\text{attn}}, \lambda_{\text{hidn}} > 0$.

During stage 2 (Domain-Adaptive Distillation)

- Student A: hard MLM + prediction-layer distillation, so $\lambda_{\text{pred}} > 0$.
- Student B: hard MLM + attention-only distillation, so $\lambda_{\text{attn}} > 0$.
- Student C: hard MLM + full intermediate distillation + prediction-layer distillation.

This design enables an ablation-style comparison showing how different distillation targets affect domain specialization and teacher imitation.

3 Results

4 Conclusion

References

[1] Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., Liu, Q.: TinyBERT: Distilling BERT for Natural Language Understanding (2020). https://arxiv.org/abs/1909.10351

[2] Hinton, G., Vinyals, O., Dean, J.: Distilling the Knowledge in a Neural Network (2015). https://arxiv.org/abs/1503.02531

[3] Ji, G., Zhu, Z.: Knowledge Distillation in Wide Neural Networks: Risk Bound, Data Efficiency and Imperfect Teacher (2020). https://arxiv.org/abs/2010.10090