Exploring the Boundaries of Low-Resource BERT Distillation

Moshe Wasserblat, Oren Pereg, Peter Izsak
Intel AI Lab, Petah Tikva, Israel

{moshe.wasserblat, oren.pereg, peter.izsak}@intel.com

Abstract

In recent years, large pre-trained models have demonstrated state-of-the-art performance in many NLP tasks. However, the deployment of these models on devices with limited resources is challenging due to the models' large computational consumption and memory requirements. Moreover, the need for a considerable amount of labeled training data also hinders real-world deployment scenarios. Model distillation has shown promising results for reducing model size, computational load and data efficiency. In this paper we test the boundaries of BERT model distillation in terms of model compression, inference efficiency and data scarcity. We show that classification tasks that require the capturing of general lexical semantics can be successfully distilled by very simple and efficient models and require relatively small amount of labeled training data. We also show that the distillation of large pre-trained models is more effective in real-life scenarios where limited amounts of labeled training are available.

1 Introduction

In recent years, large pre-trained models such as BERT (Devlin et al., 2019), GPT-2 (Radford et al., 2018) and XLNET (Yang et al., 2019) have demonstrated state-of-the-art performance in many NLP tasks and have become standard. However, the deployment of these models on devices with limited resources is challenging due to the models' large computational consumption and memory requirements. For example, the two variants of BERT, named $\text{BERT}_{\text{BASE}}$ and $\text{BERT}_{\text{LARGE}}$ consist of approximately 110M and 340M parameters, respectively. Another deployment hurdle in real-world scenarios is the scarcity of labeled data resources.

Model distillation (Ba and Caruana, 2014; Hinton et al., 2015) has shown promising results for reducing model size and computational load while preserving much of the original model's performance. A typical model distillation setup includes two stages; in the first stage, a large, cumbersome and accurate *teacher* neural network is trained for a specific downstream task. In the second stage a smaller and simpler *student* model, that is more practical for deployment in environments with limited resources, is trained to mimic the behavior of the teacher model.

Prior work related to transformer-based model distillation, focused on reducing the number of layers of the original model, obtaining shallower and more efficient student models (Sun et al., 2019; Sanh, 2019; Turc et al., 2019). Tang et al. (2019) proposed a BERT distillation method for single sentence classification tasks and sentence matching tasks using a BiLSTM (Graves, 2012; İrsoy and Cardie, 2014) student model. Our work is closely related to the work of Tang et al. (2019), however, in our work we push the boundaries of BERT model distillation in terms of model size and complexity reduction, computational load and data scarcity for single-sentence classification tasks.

The contribution of this paper is twofold; first, we show that classification tasks that require the capturing of general lexical semantics can be successfully distilled by simple and efficient models while preserving results comparable to those achieved by BERT. Second, building on previous work (Izsak et al., 2019; Mukherjee and Awadallah, 2020), we show that the distillation of large pre-trained models is more effective in real-life scenarios, where a limited amount of labeled training is available. Moreover, we show that in low data resource scenarios, the distillation model size and complexity can be substantially reduced. Specifically, we show that results produced by using a very simple and efficient model such as Continuous Bag of Words (CBoW) with a Feed Forward Network(FFN) are comparable to results produced by using a more complex model such as BiLSTM.

Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing, pages 35–40
Online, November 20, 2020.

2 Approach

The aim of a model distillation process is to use a large pre-trained *teacher* model to train a small and computationally efficient *student* model so it achieves accuracy comparable to that of the teacher model. In this section we describe the teacher and student model architectures (Sections 2.1) and the distillation process (Section 2.2).

2.1 Models Architecture

For the teacher model we chose the popular pre-trained BERT model (Devlin et al., 2019). Specifically, we used $\text{BERT}_{\text{BASE}}$, consisting of 110M parameters, and added a sentence-level softmax classification layer on top of BERT's CLS token output. The first step of the distillation process is to fine-tune BERT for a specific task using labeled data. In this step, we jointly fine-tune the parameters of BERT and the sentence-level classifier by maximizing the probability of the correct label, using the cross-entropy loss.

For student models we chose two non-transformer-based models whose neural architectures are shallower than BERT, and which contain considerably fewer parameters. The two student models are:

CBoW-FFN This simple student model is often used for very efficient text classification tasks based on sentence representation (Agibetov et al., 2018; Chen et al., 2018). The network consists of an internal embedding layer with embedding vectors of dimension $d_{emb} = 16$, followed by an average pooling layer and a Feed-Forward Network (FFN). The model contains approximately 80K parameters, meaning it is approximately 1375 times more compact than $\text{BERT}_{\text{BASE}}$.

BiLSTM The BiLSTM network (Graves, 2012; İrsoy and Cardie, 2014) consists of a pre-trained embedding[1] layer followed by two identical BiLSTM layers stacked one on top of another, and where the last hidden state of the second layer is followed by a FFN. The model contains approximately 685K parameters, meaning it is approximately 160 times more compact than $\text{BERT}_{\text{BASE}}$.

Additional Models We also experimented with Convolutional Neural Networks (CNNs) (Kalchbrenner et al., 2014). However, BiLSTM performed better for the same model size.

[1] We used Stanford GloVe embeddings https://nlp.stanford.edu/projects/glove/

Dataset	Task	T-train	S-train	Test
AGNews	topic	400	20K	7.6K
Emotion	emotion	1000	50K	2K
IMDB	sentiment	1000	25K	25K
SST-2	sentiment	200	1M*	1.9K
CoLA	acceptability	1000	1M*	516

Table 1: Dataset descriptions and statistics. T-train represents the number of labeled samples used for training the teacher model (step 1) and S-train represent the number of unlabeled samples used for training the student model (step 2). *Obtained using the data augmentation method described by Jiao et al. (2020).

2.2 The Distillation Process

The first step of the training process consists of fine-tuning the teacher model using the available labeled data. The second step of the distillation process is depicted in Figure 1. In this step the student model is trained using the unlabeled data. The unlabeled data is fed in parallel into both the fine-tuned teacher model and to the student model. Following (Tang et al., 2019), we only use the distillation loss which is calculated for each training batch by performing Mean Square Error (MSE) between the soft targets (logits) that are produced by the student and teacher models:

$$L_{distill} = \frac{1}{N}\sum_{n=0}^{N}(y_s - y_t)^2$$

where y_s and y_t are the logits produced by the student and teacher models, respectively.

3 Experimental Setup

3.1 Datasets and Tasks

The goal of our work is to test the distillation boundaries in terms of model size compression, inference computation load and training data size of single-sentence classification tasks. We conducted experiments on five widely-used single-sentence classification datasets, as detailed below.

AGNews A topic classification dataset (Zhang et al., 2015) that consists of internet news titles labeled with four categories: World, Entertainment, Sports and Business.

Emotion An emotion classification dataset (Saravia et al., 2018) that consists of Twitter posts labeled with any of six basic emotion categories: sadness, disgust, anger, joy, surprise, and fear.

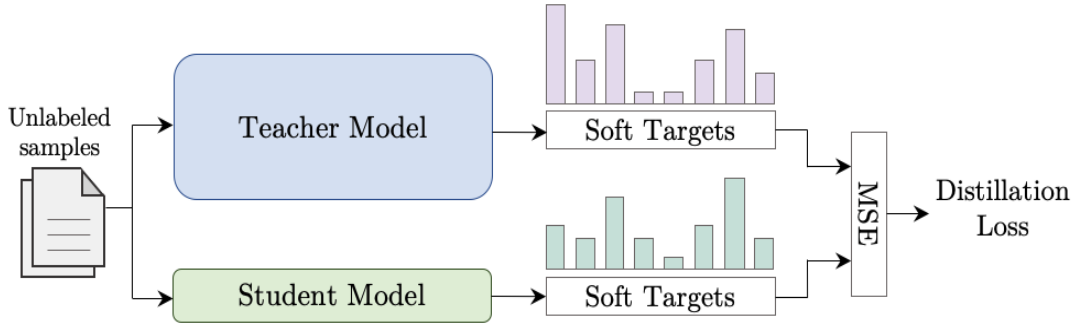

Figure 1: Student model training process. The student and teacher models process the unlabeled samples and generate logits for each example. Distillation loss is produced by calculating the Mean Square Error(MSE) between the logits of both models.

Model	AGnews	Emotion	IMDB	SST-2	CoLA	Comp. ratio	Speedup
BERT_{BASE}	87.3	82	88.3	83.5	56	x1	x1
CBoW-FFN	86.3	82	87.6	79.1	10	x1375	x574
BiLSTM	86.4	81.8	85.6	80.7	10	x160	x40

Table 2: Low-data-resource distillation models comparison. For all datasets we report the F1 score except for CoLA, for which we report the Matthews Correlation Coefficient (MCC). Comp. ratio and Speedup[2] represent the model size reduction ratio and inference speedup, respectively, in relation to BERT_{BASE}.

IMDB The Internet Movie Database (IMDB; Maas et al. 2011) comprises single sentences extracted from informal movie reviews for binary (positive/negative) sentiment classification.

SST-2 The Stanford Sentiment Treebank 2 (SST-2; Socher et al. 2013) comprises single sentences extracted from movie reviews for binary (positive/negative) sentiment classification. This dataset is part of the widely used General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2018).

CoLA The Corpus of Linguistic Acceptability (CoLA; Warstadt et al. 2018) consists of English acceptability judgments drawn from books and journal articles on linguistic theory. Each sentence is annotated with whether it is a grammatical English sentence or not. This dataset is also part of the GLUE benchmark.

Table 1 shows the dataset descriptions and statistics. In order to simulate a real-life data-scarce environment, we limited the labeled teacher model training set (T-train) size to no more than a thousand samples. It was shown that large amounts of data are needed for the teacher model to fully express its knowledge (Ba and Caruana, 2014). For AGNews, Emotion and IMDB datasets, we used the available training data which is part of the datasets as unlabeled student training data (S-train). However, both SST-2 and CoLA datasets, do not contain sufficient amounts of training data, therefore, we use the data augmentation method described by Jiao et al. (2020) for generating unlabeled student training data (S-train).

3.2 Setup

We adopt the HuggingFace (Wolf et al., 2019) implementation of BERT-base (uncased)[3] model for the teacher model. We fine-tune the model for 3 epochs with learning rate of $5e^{-5}$ and batch size of 16. The CBoW-FFN student model was implemented based on the model described by Agibetov et al. (2018) with embedding size of 16 and word vocabulary size of 5000. The BiLSTM student model was implemented in a fashion similar to the model described by Chollet[4] with embedding size of 100 and with vocabulary size of 5000.

4 Results and Discussion

4.1 The Low Resource Scenario

Table 2 shows low-data-resource scenario comparison between the accuracy of the two student

[2]Runnig on Intel(R) Xeon(R) CPU @ 2.30GHz, OS: Ubunto 18.04.3 LTS and Tensorflow 2.2

[3]https://github.com/huggingface/transformers

[4]https://keras.io/examples/nlp/bidirectional_lstm_imdb/

Model	SST2-low resource*	SST2-high resource**
BERT_{BASE}	83.5	91
CBoW-FFN	79.1	82
BiLSTM	80.7	86.1
CBoW-FFN-NoDs†	62.8	81.2
BiLSTM-NoDs†	63.1	78.8

Table 3: F1 score comparison between low and high data resource scenarios for the SST-2 dataset. *Teacher model training size = 200 samples. **Teacher model training size = 6920 samples. †No distillation.

models and the teacher model across the different datasets and tasks. Overall, the distilled models produced results that are competitive with the teacher model's results across all datasets and tasks except for the CoLA task. An interesting observation is that the relatively lightweight CBoW-FFN model's results are on-par with the BiLSTM results. A possible explanation for these results is that all of the tasks, with the exception of CoLA, require the detection of general lexical semantic features with relatively less emphasis on linguistic structure and contextual relations, therefore BERT's contextual-oriented architecture has no advantage over the student models' architecture. The CoLA task, on the other hand, requires the detection of linguistic structure and contextual relations and this is where BERT's architecture excels and the student models' architectures are lacking.

4.2 Low Resource Vs. High Resource

Table 3 shows an F1 score comparison between the two student models and the teacher model for low and high labeled data resource scenarios for the SST-2 dataset. The table also shows results for the student models when trained directly on the labeled data (non-distilled version).

Distilled Vs. Non-Distilled Models The results demonstrate that the student models trained using the distillation method (described in Section 2.2), consistently outperform the baseline student models trained directly on the labeled data, proving the effectiveness of the distillation approach. However, and in accordance with the findings of Izsak et al. (2019); Mukherjee and Awadallah (2020), it is also evident that the F1 score enhancement achieved by the distilled student models over the non-distilled models is higher in the low resource scenario than in the high resource scenario. Specifically, the F1 improvement between the distilled and non-distilled versions of the two student models in the low resource scenario are 16.3% and 17.6%, vs. 0.8% and 7.3% in the high resource scenario.

Distilled Models Vs. BERT The results also show that in the high resource scenario case, where an abundance of labeled training data is available, BERT's accuracy advantage over the distilled models grows larger compared to the low-resource scenario. Specifically, the F1 score gaps between BERT and the student models in the high resource scenario are 9%, and 4.9%, respectively, whereas in the low resource scenario those gaps are only 4.4% and 2.8% respectively.

BiLSTM Vs. CBoW-FFN Another observation is that in the high resource case, the practical trade-off between model complexity and accuracy becomes more salient. For example, the F1 score gap between CBoW-FFN and BiLSTM is merely 1.6% in the low resource scenario but reaches 4.1% in the high resource scenario. This observation aligns with the basic neural-networks phenomena that larger and deeper neural networks are able to represent the distribution of the data more accurately compared to smaller models when large amounts of data are available (Ng, 2018).

Practical Implications The practical implications of these results is that distillation is more effective in real-life scenarios where limited amounts of labeled training data are available. In high-resource scenarios, however, where an abundance of labeled training data is available, using deeper and more complex student models such as BiLSTM, or shallower transformer-based models, yields higher accuracies.

5 Conclusion

We showed that in low resource scenarios, it is feasible to distil BERT using very efficient models while preserving comparable results. However, the success of the distillation depends on the dataset and task at hand. Classification tasks that require capturing of general lexical semantics can be successfully distilled by very simple and efficient models; however, classification tasks that require the detection of linguistic structure and contextual relations are more challenging for distillation using simple student models. For future work, we aim to explore the impact of the datasets' linguistic structures on the distillation success and to develop dataset-related measurements (Arora et al., 2020) for predicting the success of the distillation in relation to different student models.

References

Asan Agibetov, Kathrin Blagec, Hong Xu, and Matthias Samwald. 2018. Fast and scalable neural embedding models for biomedical sentence classification. BMC Bioinformatics 19, 541.

Simran Arora, Avner May, Jian Zhang, and Christopher Ré. 2020. Contextual embeddings: When are they worth it? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2650–2663, Online. Association for Computational Linguistics.

Jimmy Ba and Rich Caruana. 2014. Do deep nets really need to be deep? In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2654–2662. Curran Associates, Inc.

Qingyu Chen, Yifan Peng, and Zhiyong lu. 2018. Biosentvec: creating sentence embeddings for biomedical texts.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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. Association for Computational Linguistics.

Alex Graves. 2012. *Supervised Sequence Labelling with Recurrent Neural Networks*. Studies in Computational Intelligence. Springer, Berlin.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. Cite arxiv:1503.02531 Comment: NIPS 2014 Deep Learning Workshop.

Ozan İrsoy and Claire Cardie. 2014. Opinion mining with deep recurrent neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 720–728, Doha, Qatar. Association for Computational Linguistics.

Peter Izsak, Shira Guskin, and Moshe Wasserblat. 2019. Training compact models for low resource entity tagging using pre-trained language models. *ArXiv*.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. Tiny{bert}: Distilling {bert} for natural language understanding.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 655–665, Baltimore, Maryland. Association for Computational Linguistics.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 142–150, Stroudsburg, PA, USA. Association for Computational Linguistics.

Subhabrata Mukherjee and Ahmed Hassan Awadallah. 2020. Distilling bert into simple neural networks with unlabeled transfer data. *ArXiv*.

Andrew Ng. 2018. *Machine Learning Yearning*, pages 11–12. deeplearning.ai.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018. Language models are unsupervised multitask learners.

Victor Sanh. 2019. Introducing distilbert, a distilled version of bert. Medium.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019. Patient knowledge distillation for BERT model compression. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4323–4332, Hong Kong, China. Association for Computational Linguistics.

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

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962v2*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. 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. Association for Computational Linguistics.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2018. Neural network acceptability judgments. *CoRR*, abs/1805.12471.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. Cite arxiv:1906.08237Comment: Pretrained models and code are available at https://github.com/zihangdai/xlnet.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 649–657. Curran Associates, Inc.