Model Distillation for Sentiment Analysis

Jiali Lu
jialilu.physics@gmail.com

***Abstract*—In this project, I consider model distillation when the data labels are scarce but unlabelled data is plentiful. This is not covered in academic studies but is often the case in industry setting. Results suggest that model distillation offers substantial gain in accuracy even in such scenario.**

1. Introduction

Model distillation (also known as knowledge distillation) is an important technique in natural language processing (NLP). In this project, I demonstrate my deep learning skills by performing model distillation for a sentiment analysis task. In addition, I try to answer the following engineering question: is model distillation still beneficial when labelled dataset is tiny but unlabelled data is plentiful, as is often the case when we start to build a new functionality.

2. Result

Four BERT based models are trained under different setup to perform sentiment analysis task on Yelp reviews dataset [1]. The training hyper-parameters are selected following recommendation of [2]. The models are:

1) BERT-base fine-tuned on 10,000 labelled reviews;
2) BERT-small distilled from BERT-base above, using 639,000 reviews without ground truth label;
3) BERT-small directly fine-tuned on the same 10,000 labelled reviews as in 1);
4) BERT-small directly fine-tuned on 20,000 labelled reviews, which includes those in 1) and 10,000 more new reviews. The extra training cost for model distillation can be used to annotate 10,000 more reviews.

The results are summarized in the table below. We can make the following observation:

Having BERT-base as a teacher increases accuracy for the BERT-small model. This is likely because larger model is less prone to get stuck at a bad local optima, and the teacher model can guide the student towards a better local optima.

The distilled BERT-small model is about 70% smaller and has 7 times faster inference speed. This makes it better suited for production.

It is surprising to find the BERT-small student model has slightly higher accuracy than the BERT-base teacher model. In this specific case, it is likely because the sample size is very small and BERT-base may have already overfitted certain rare patterns. However, the overfitting was not passed on since the student model is much smaller and the rare patterns may not appear a lot in the larger dataset for distillation.

3. Conclusion

Model distillation offers accuracy close to that of the large model, while featuring the size and speed of a smaller model. It remains useful even when the labelled dataset is tiny. Thus, it is a valuable tool even in early stage of development.

References

[1] Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. Advances in neural information processing systems, 28.

[2] Mosbach, M., Andriushchenko, M., & Klakow, D. (2020). On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. arXiv preprint arXiv:2006.04884.

TABLE 1. SUMMARY OF FOUR BERT BASED MODELS

	model	training procedure	sample size	training time (h)	accuracy (%, polarity)	inference speed (sample/s)
1	BERT-base	fine-tuning	10k	4.05	94.9	48.7
2	BERT-small	model distillation	10k / 639k *	12.2	95.6	337.8
3	BERT-small	fine-tuning	10k	0.58	93.5	335.9
4	BERT-small	fine-tuning	20k	1.18	94.2	339.4

*: trained on 639k reviews soft-labelled by BERT-base above