Learning small sentence classifiers from BERT using teacher-student knowledge distillation

Sam Sučík

MInf Project (Part 2) Report
Master of Informatics
School of Informatics
University of Edinburgh

2020

Abstract

Acknowledgements

Thanks to Steve and Vova, to Ralph Tang and to Slávka.

Table of Contents

Chapter 1

Introduction

1.1 Motivation

- After the deep learning hype started, NLP went through an era of LSTMs. Since 2017, the area has been becoming dominated by Transformer models pre-trained on large unlabelled corpora.

- As newer and bigger Transformer-based models were proposed in 2018 and 2019, improving on the SOTA, it was becoming clearer that their big size and low speed was rendering them difficult to use (both train and deploy) in practice outside of research labs.

- Recently, we've seen various early attempts at making Transformers – in particular BERT (Devlin et al., 2018) – smaller by removing attentional heads (Michel et al., 2019), quantisation and pruning (Cheong and Daniel, 2019; Sucik, 2019). In terms of actually down-sizing and accelerating the models, knowledge transfer using teacher-student knowledge distillation has led to the most attractive results (Mukherjee and Awadallah, 2019; Tang et al., 2019; Jiao et al., 2019; Sanh et al., 2019).

- However, these studies focus only on using knowledge distillation as a tool. Important questions about the nature of this technique and how it interacts with properties of the teacher and student models remain generally unexplored.

- In line with the increasind demand for explainable AI, it is desirable to better understand how knowledge distillation works, in this case when distilling NLP knowledge from Transformer models. Such understanding can also help researchers improve the ways knowledge distillation is used or modified for various use cases.

1.2 Aims

- Explore the effectiveness of knowledge distillation in very different NLP tasks. To cover a broad variety of tasks, I use datasets ranging from binary sentiment classification to 57-way intent classification to linguistic acceptability.
- Explore how distilling knowledge from a Transformer varies with different student architectures. I limit myself to using the extremely popular BERT model (Devlin et al., 2018) as the teacher architecture. As students, I use two different architectures: a BiLSTM, building on the successful work of Ralph Tang (Tang et al., 2019; Tang and Lin, 2019), and a down-scaled BERT architecture.
- Explore how successfully can different types of NLP knowledge and capabilities be distilled. Since NLP tasks are often possible for humans to reason about, I analyse the models' behaviour (e.g. the mistakes they make) to learn more about knowledge distillation. I also probe the models for different linguistical capabilities, building on previous succesful probing studies (Conneau et al., 2018; Tenney et al., 2019a).

1.3 Contributions

My actual findings. To be added later.

Chapter 2

Background

2.1 Brief history of pre-Transformer sequence encoders

NLP is all about sequences of variable lengths: sentences, sentence pairs, documents, speech segments...

NLP tasks are typically about making simple predictions about sequences: classifying sentences based on their intent or language, scoring a document's level of formality, predicting whether two sentences form a coherent question-answer pair or not, predicting the next word of a sentence...

Machine learning predictors are typically designed to work with fixed-size representations of inputs. Therefore, ever since the resurgence of neural networks around 2010, neural NLP has been using various models for encoding variable-length sequences into common fixed-dimensional representations.

RNN- and later LSTM-based encoder architectures were dominating the area for a long time as they were naturally suited for processing sequences of any length.

A major breakthrough came when Kalchbrenner and Blunsom (2013) and Sutskever et al. (2014) developed the encoder-decoder architecture for machine translation and other sequence-to-sequence tasks such as paraphrasing or parsing.

Bahdanau et al. (2014) improved things by introducing attention, enabling the recurrent encoders to learn to selectively attend or ignore parts of the input sequence.

2.2 Transformer models

Vaswani et al. (2017) introduced Transformer. Radford et al. (2018) introduced the idea of generative pre-training and fine-tuning. Devlin et al. (2018) introduced BERT (i.e. bi-directionality and 2 simultaneous pre-training tasks – MLM and NSP). further huge models include: GPT-2 (Radford et al., 2019), XLM (Lample and Conneau, 2019) (introduced cross-lingual pre-training).

2.3 Knowledge distillation

2.4 Understanding the models

Chapter 3

Datasets

Chapter 4

Implementation

Chapter 5

Student tuning

Chapter 6

Analysing the students

Chapter 7

Overall discussion and conclusions

Chapter 8

Future work

Chapter 9

Notes

9.1 Distillation techniques

Papamakarios (2015, p. 13) point out that mimicking teacher outputs (e.g. with cross-entropy loss) can be taken to next level by mimicking the derivatives of the loss w.r.t. inputs (i.e. including in the KD loss function also this term: $\frac{\partial \mathbf{o}_{student}}{\partial \mathbf{x}} - \frac{\partial \mathbf{o}_{teacher}}{\partial \mathbf{x}}$), the additional loss term being calculated using the R technique (Pearlmutter, 1994).

Sau and Balasubramanian (2016) show that learning from noisy logits helps (adding the noise is very simple).

Kim and Rush (2016) observe that KD and weight pruning are orthogonal (can be used together), and that mimicking top-most hidden layer outputs (instead of outputs themselves) doesn't provide improvements previously reported.

Huang and Wang (2017) propose method for matching neuron activation distributions of teacher and student (only suitable for same teacher/student architecture?). Heo et al. (2018) do a similar thing but try to match the activation boundaries of neurons.

9.2 Knowledge distillation – assorted

Zharov et al. (2018) use KD of DNNs into decision forests for interpretability.

Mirzadeh et al. (2019) show that a large teacher cannot teach too small students, and that adding intermediate ”teacher assistants” helps.

9.3 Datasets

Unsuitable:

1. ATIS: too easy (see here). Rasa version here.

2. AskUbuntu, Chatbot and Web Applications (all from TU Munich): too small (max. 206 datapoints). Rasa version here.
3. SNIPS: too easy (see here). Rasa version here.

Suitable:

1. FB's Multilingual Task Oriented Dataset: F1 0.99 by supervised embeddings (very easy, but perhaps usable). Rasa version here.
2. CoLA (Warstadt et al., 2018), needs processing into Rasa format.
3. SST (Socher et al., 2013), needs processing into Rasa format. 5-way classification may be too hard (accuracy ~50%), 2-way much easier.
4. TREC question-type classification (Voorhees and Tice, 2000), needs processing into Rasa format. 6-way classification (abbreviation, entity, description, human, location, numeric), also has another (more fine-grained) level of categories.

Maybe suitable:

1. Microsoft Dialogue Challenge (Li et al., 2018) needs processing into Rasa format. Also, no results using Rasa.
2. TOP (Gupta et al., 2018) needs processing into Rasa format. Also, no results using Rasa. Intents are hierarchical, would need to take only the top-most intent.
3. SciCite (Cohan et al., 2019) needs processing into Rasa format. Also, no results using Rasa.

Probing/evaluation: Started by Shi et al. (2016) and Adi et al. (2017)?

1. Google's edge probing (Tenney et al., 2019b) for evaluating span representations on 9 tasks closely following classical NLP pipeline (data not freely accessible!)
2. FB's probing (Conneau et al., 2018) used to evaluate entire sentence embeddings (10 tasks from sentence length to semantics)
3. FB's SentEval (Conneau and Kiela, 2018) is meant for evaluating trained sentence encoders (i.e. not meant as downstream tasks that encoders should be fitted to), but it curates interesting existing datasets.

9.4 Plans for MVP

Let's distill BERT into a smaller BERT.

Code: pytorch-transformers (re-using code for DistilBERT), no Rasa for now.

Dataset: CoLA (because pytorch-transformers offers easy feeding of GLUE datasets).

Bibliography

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate.

Cheong, R. and Daniel, R. (2019). transformers.zip: Compressing transformers with pruning and quantization.

Cohan, A., Ammar, W., van Zuylen, M., and Cady, F. (2019). Structural scaffolds for citation intent classification in scientific publications. In *NAACL-HLT*.

Conneau, A. and Kiela, D. (2018). Senteval: An evaluation toolkit for universal sentence representations. *arXiv*, abs/1803.05449.

Conneau, A., Kruszewski, G., Lample, G., Barrault, L., and Baroni, M. (2018). What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *ACL*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding.

Gupta, S., Shah, R., Mohit, M., Kumar, A., and Lewis, M. (2018). Semantic parsing for task oriented dialog using hierarchical representations. In *EMNLP*.

Heo, B., Lee, M., Yun, S., and Choi, J. Y. (2018). Knowledge transfer via distillation of activation boundaries formed by hidden neurons. *CoRR*, abs/1811.03233.

Huang, Z. and Wang, N. (2017). Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv*, abs/1707.01219.

Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., and Liu, Q. (2019). TinyBERT: Distilling BERT for natural language understanding.

Kalchbrenner, N. and Blunsom, P. (2013). Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709.

Kim, Y. and Rush, A. M. (2016). Sequence-level knowledge distillation. *arXiv*, abs/1606.07947.

Lample, G. and Conneau, A. (2019). Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Li, X., Panda, S., Liu, J., and Gao, J. (2018). Microsoft dialogue challenge: Building end-to-end task-completion dialogue systems. *arXiv*, abs/1807.11125.

Michel, P., Levy, O., and Neubig, G. (2019). Are sixteen heads really better than one?

Mirzadeh, S., Farajtabar, M., Li, A., and Ghasemzadeh, H. (2019). Improved knowledge distillation via teacher assistant: Bridging the gap between student and teacher. *CoRR*, abs/1902.03393.

Mukherjee, S. and Awadallah, A. H. (2019). Distilling transformers into simple neural networks with unlabeled transfer data.

Papamakarios, G. (2015). Distilling model knowledge (MSc thesis). *arXiv*, abs/1510.02437v1.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.

Sau, B. B. and Balasubramanian, V. N. (2016). Deep model compression: Distilling knowledge from noisy teachers. *arXiv*, abs/1610.09650.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*.

Sucik, S. (2019). Pruning BERT to accelerate inference. https://blog.rasa.com/pruning-bert-to-accelerate-inference/.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Tang, R. and Lin, J. (2019). Natural language generation for effective knowledge distillation.

Tang, R., Lu, Y., Liu, L., Mou, L., Vechtomova, O., and Lin, J. (2019). Distilling task-specific knowledge from BERT into simple neural networks. *arXiv*, abs/1903.12136.

Tenney, I., Das, D., and Pavlick, E. (2019a). BERT rediscovers the classical NLP pipeline. In *ACL*.

Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., Kim, N., Durme, B. V., Bowman, S. R., Das, D., and Pavlick, E. (2019b). What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv*, abs/1905.06316.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Voorhees, E. M. and Tice, D. M. (2000). Building a question answering test collection. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, pages 200–207, New York, NY, USA. ACM.

Warstadt, A., Singh, A., and Bowman, S. R. (2018). Neural network acceptability judgments. *arXiv*, abs/1805.12471.

Zharov, Y., Korzhenkov, D., Shvechikov, P., and Tuzhilin, A. (2018). YASENN: explaining neural networks via partitioning activation sequences. *arXiv*, abs/1811.02783.