# NLP knowledge distillation with different students

*Sam Sučík*

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

2020

# Abstract

# Acknowledgements

# Table of Contents

# Chapter 1

# Notes

## 1.1 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.

## 1.2 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.

## 1.3 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:

1. Google's edge probing (Tenney et al., 2019) 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.

## 1.4 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


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*.

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.

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

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.

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.

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

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*.

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. (2019). What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv*, abs/1905.06316.

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.