Faculté des arts et des sciences
Département d'**informatique et de recherche opérationnelle**

Pretrained Language Models on Low Ressources

Gao Yinghan, Wei Wei, Frederic Boileau

February 5, 2021

Transformer Vaswani et al [1] introduced a revolutionnary architecture for machine translation in 2017, the transformer. The idea was based on the idea of an encoder and decoder for transduction as for the state of the art predecessors relying on deep bidirectionnal RNNs for both the encoding and the decoding stages. The authors of the transformer paper, "Attention is all you need" based their architecture's sequence modeling solely on an attention mechanism which enables it to model the long range dependencies in a sentence without the inherently sequential constraint which RNNs in their various forms imply. The massive parallelization enabled massive improvements in training time to achieve state of the art results in machine translation amongst other NLP tasks.

Pretrained Language Models Bidirectional Encoder Representation from (BERT)[2] is an architecture based on the encoder module of the transformer. Its training is done in two steps, first it learns a Language Model (LM) which results in a Pretrained Language Model (PLM) through some unsupervised learning tasks. The training procedure is done over two tasks: Masked Language Model (MLM) and Next Sentence Prediction (NSP). The MLM task simply trains the model to learn to predict randomly masked words from a sentence. In NSP, BERT learns to predict, given a pair of sentences, whether one follows the other. When trained on those tasks the BERT yields a PLM. The resulting model can then be fine-tuned in a supervised setting for a specific tasks such as question answering. This ability to train once over a huge corpus of data to yield a LM and then fine tune it downstream adresses one core issue of deep learning: how to transfer learning or knowledge.

BERT is just one of many examples of large PLM deployed today. The following table lists the models and their associated number of parameters.

Architecture	Number of parameters
BERT	340M
GPT-2	1.5B
MegatronLM	8.3B
T5	11B
T-NLG	17B
GShard	600B

Table 1: PLMs and their sizes[3]

Knowledge Distillation Knowledge Distillation (KD) addresse the following issue, how can we leverage the state of the art (SOA) results given by large PLMs to do inference in a context where memory and computing power are limited. One avenue would be to use one of the large PLMs to “teach” a smaller model (the student) which we can deploy in more ressource limited environments. Hinton, though not specifically with respect argued in 2015[4] that one conceptual roadblock to knowledge transfer or “distillation” had been the rigid identification of a model with the learned parameter values instead of the more abstract view of a “learned mapping from input vectors to output vectors”[4]. The way to do this according to Hinton et al is to make the student learn through an objective function which reflects the generalization ability learned in the teacher model. To achieve this he proposes using an objective function which averages over the soft target (the output probabilities of the teacher where the logits are divided by temperature factor to adjust the smoothness of those targets) and the ground truth.

TinyBERT ...[**?**]

Further model compression We have the following other model compression techniques...[3]

- Pruning...
- Quantization...
- Parameter Sharing...
- Tensor Decomposition...

Evaluation

- Squad[5]
- GLUE[6]

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[3] Manish Gupta and Puneet Agrawal. Compression of deep learning models for text: A survey, 2020.

[4] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

[5] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don’t know: Unanswerable questions for squad. *CoRR*, abs/1806.03822, 2018.

[6] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. SWAG: A large-scale adversarial dataset for grounded commonsense inference. *CoRR*, abs/1808.05326, 2018.