Université de Montréal

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

The Transformer revolution

Vaswani et al [1] introduced a revolutionnary architecture in 2017, the transformer. Still based on the idea of an encoder and decoder for transduction, they based their architecture solely on an attention mechanism enables 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

Pretrained language models and BERT

BERT is ...[2]
PLMs in general are ...
But those models tend to be huge ...

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]

Compression approaches

Model compression consists of ...

Knowledge distillation

Knowledge Distillation (KD)...[4]

tinyBERT

TinyBERT...[5]

Other model commpression techniques

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

Pruning ...

Quantization ...

Parameter Sharing ...

Tensor Decomposition ...

Experiments and Evaluation

PLMs can be evaluated on GLUE for general LM performance and more specific downstream tasks for which they are fine tuned such as Squad.

SquAD ...[6]

GLUE ...[7]

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] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling BERT for natural language understanding. *CoRR*, abs/1909.10351, 2019.

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

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