Compressing BERT - An Evaluation and Combination of Methods

Sebastian Jäger

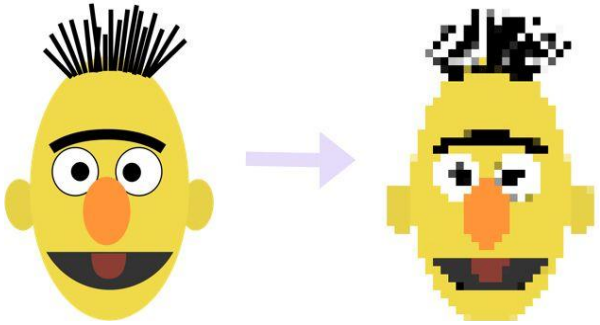
Model compression aims to reduce the number of parameters without or with as little performance loss as possible.

Cristian Bucila, Rich Caruana, and Alexandru Niculescu-Mizil. “Model Compression”. In: Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). 2006, pp. 535–541. doi: 10.1145/1150402.1150464.

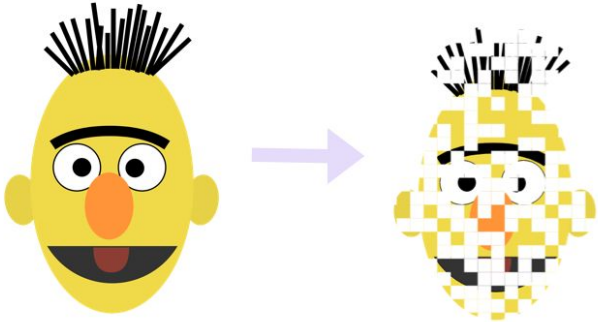
T5
11,000

GPT-3
175,000

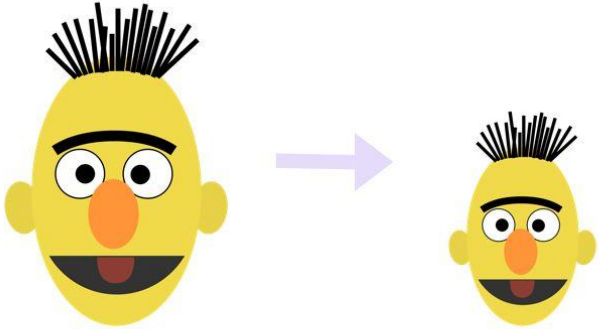
Quantization



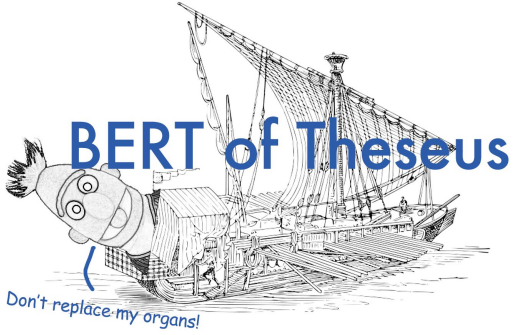
Pruning



Knowledge Distillation



Theseus Compression



Images: https://blog.rasa.com/compressing-bert-for-faster-prediction-2/
https://raw.githubusercontent.com/JetRunner/BERT-of-Theseus/master/bert-of-theseus.png

“Using a domain-specific BERT compressed with Theseus Compression helps to speed up experiment iterations for downstream task models.”

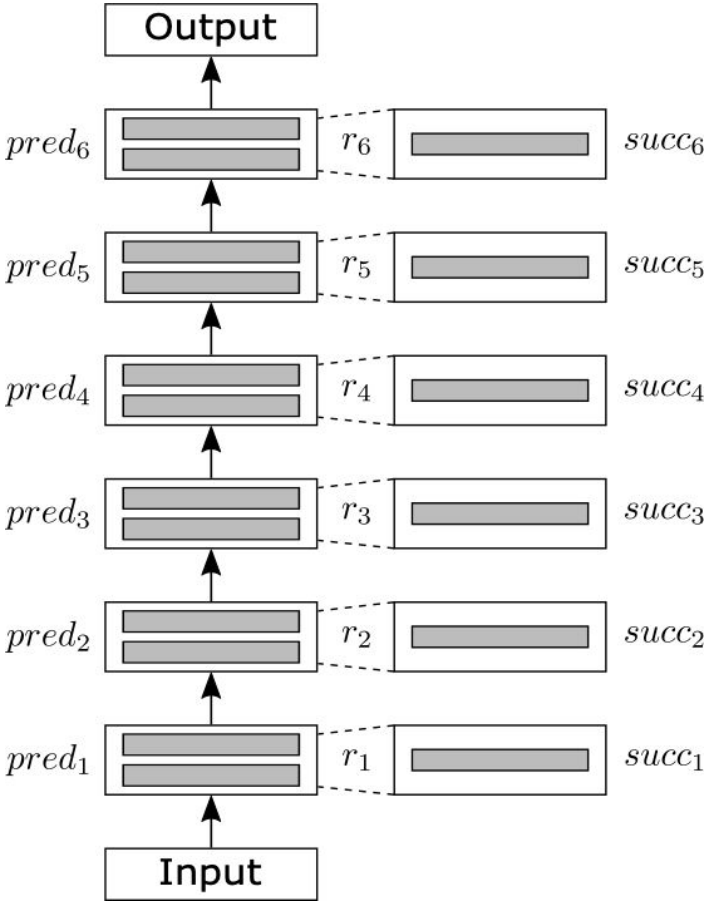
Tasks and Datasets

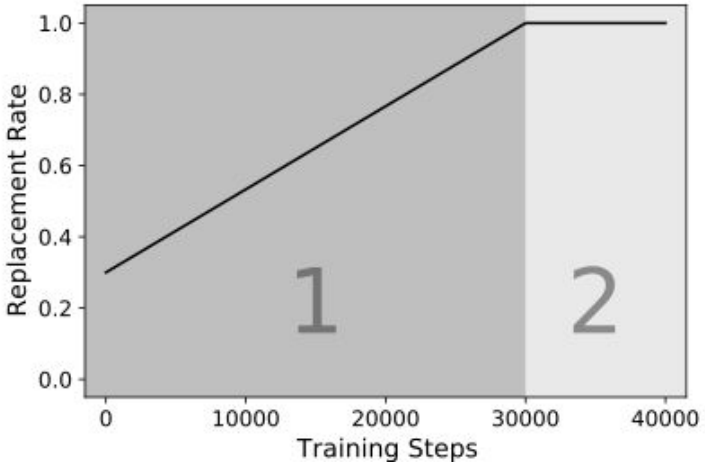
In-hospital Mortality Prediction

- MIMIC-III

German Hate Speech Detection

- GermEval 19
- NOHATE + NOHATE LM


Output
$pred_6$
r_6
$succ_6$
$pred_5$
r_5
$succ_5$
$pred_4$
r_4
$succ_4$
$pred_3$
r_3
$succ_3$
$pred_2$
r_2
$succ_2$
$pred_1$
r_1
$succ_1$
Input


Replacement Rate
1.0
0.8
0.6
0.4
0.2
0.0
1
2
0
10000
20000
30000
40000
Training Steps

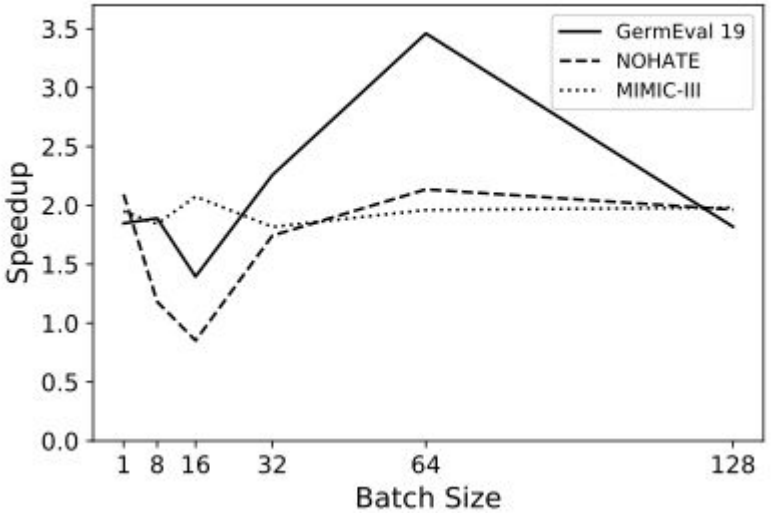
Results of Hyperparameter Analysis

- Starting replacement rate around 0.1 to 0.3
- Module replacement for 85% of compression steps
- Both module replacement and successor fine-tuning are important
- Learning rate plays minor role in the compression performance

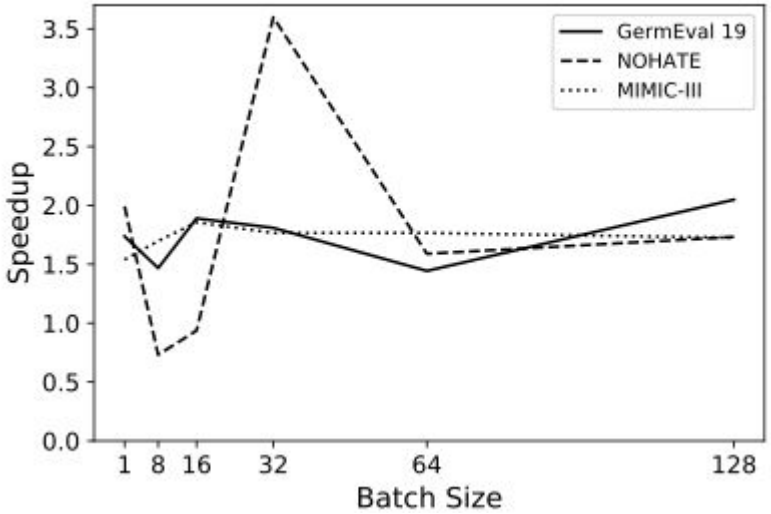
Model Settings	Retained Performance	Speedup Fine-tuning
NOHATE LM + GermEval 19	0.9750	2.50×
NOHATE LM + NOHATE	0.9889	2.16×
MIMIC-III LM + MIMIC-III	0.9947	2.14×

Prediction Speedup

CPU



GPU



Qualitative Error Analysis

100 FPs and FNs for baseline and compressed models.

- **GermEval 19:** Change of -5 and introduced 2 new error classes
- **NOHATE:** More stable, mostly changes within ± 1, max +3
- **MIMIC-III:** All changes within ± 1

Summary of Results

- Compression faster than plain LM fine-tuning
- Compression ratio: 1.67×
- Retained performance: up to 99%
- Speedup downstream training: >2×
- Speedup inference: 1.94× on CPU, 1.73× on GPU

Conclusion and Limitation

+ Theseus Compression speeds up experiments
+ Compressed models are faster and smaller

- Theseus Compression always reduces prediction performance

Thank You!

Questions?