FastBERT: Speeding up Self-attentions

Bodhisattwa Prasad Majumder, Huanru Henry Mao, Khalil Mrini
Computer Science and Engineering
University of California, San Diego
9500 Gilman Dr, La Jolla, CA 92093
{bmajumde,hhmao,kmrini}@eng.ucsd.edu

Proposal

Bi-directional Encoder Representations for Transformers (BERT) (Devlin et al., 2018) have achieved state-of-the-art performance in many natural language understanding tasks on the GLUE benchmark (Wang et al., 2018). The model architecture consists of 12 stacked blocks of 12 multi-head self-attentions followed by 3072 units of point-wise feed-forward network, totaling 110 million parameters. Such a large model is useful when computational power is abundant, but is too heavy to deploy on mobile devices. In this work, we aim to build a compact and less computationally expensive variant of BERT, and measure the trade-off with accuracy.

Producing a compact variant of BERT is likely feasible due to the Lottery Ticket Hypothesis (Frankle and Carbin, 2018), which suggests that only a subset of a network’s parameters are essential to achieving its accuracy during inference. The Knowledge Distillation method (Hinton et al., 2015) explores an alternative approach to compress a model’s parameters by transferring generalizations from a large teacher model to a small student model using the teacher’s “dark knowledge” (high temperature Softmax outputs). FitNets (Romero et al., 2014) build upon the Knowledge Distillation technique and provide gradient information to intermediate layers of a deep network to compress wide models into deep and thin models, achieving 13x speed up.

There has also been promising recent work on optimizing neural components within the Transformer (Vaswani et al., 2017) architecture to improve its computational and memory efficiency. Huang et al. (2018) aim to improve memory efficiency of the Music Transformer by implementing relative position-based attention, reducing the memory requirement from $O(L^2D)$ to $O(LD)$, where $L$ and $D$ are the sequence length and hidden state dimension respectively. In the Weighted Transformer (Ahmed et al., 2017), the attention layers are modified to converge 15 to 40% faster by replacing multi-head attention by a mechanism similar to mixture of experts. The Universal Transformer (Dehghani et al., 2018) demonstrates that we can do better in machine translation if we simply tie all the weights across transformer blocks, reducing the amount of model parameters. These work suggest that the Transformer is not an optimal architecture and there are room for improvements.

We propose FastBERT, a thinner yet deeper version of BERT to achieve both computational and memory efficiency for highly expensive yet very effective self-attention architecture. In this paper, we would like to adopt the work done by Romero et al. (2014) to distill knowledge from BERT. Our thinner network will be guided by the hidden layers and internal representations of the original BERT model.\footnote{Pre-trained network parameters are released by Google: https://github.com/huggingface/pytorch-pretrained-BERT}

We plan to test our proposed method on some of BERT’s text classification benchmarks from GLUE (Wang et al., 2018) such as SST-2, COLA and the Amazon Review dataset (McAuley and Leskovec, 2013). We will conduct experiments with FastBERT and measure the trade-off between speed and accuracy, in a similar way to Facebook’s fastText classifier (Joulin et al., 2016).

References


Mostafa Dehghani, Stephan Gouws, Oriol Vinyals,


