TN-BERT (TensorNetwork BERT)

TN-BERT is a modification of the BERT-base architecture that greatly compresses the original BERT model using tensor networks. The dense feedforward layers are replaced with Expand / Condense tn layers tuned to the TPU architecture.

This work is based on research conducted during the development of the TensorNetwork Library. Check it out on github.

TN-BERT achieves the following improvements:

- 69M params, or 37% fewer than the original BERT base.
- 22% faster inference than the baseline model on TPUs.
- Pre-training time under 8 hours on an 8x8 pod of TPUs.
- 15% less energy consumption by accellerators

For more information go to the TF Hub model page here

Implementation

The expand_condense and transformer layers are the only components that differ from the reference BERT implementation. These layers can be viewed at:

- tn_transformer_expand_condense.py
- tn_expand_condense.py