Appendix

A IMPLEMENTATION DETAILS

In out experiments we used `BERT-large-uncased` with English Wikipedia (2,500M words) and BookCorpus (800M words) (Zhu et al., 2015) datasets as training the data with the LAMB optimizer. First, we performed the model pre-training when applying Variance Pruning on top of the original regime suggested by Devlin et al. (2018). In the first MaskedLM phase, we start with randomly initialized model and train on a DGX-A100 cluster with 8 GPUs with 32GB of memory each. We used a batch size of 64 per device and train for 7038 training steps. We used a learning rate of 6e-3 and a maximum sequence length of 128. Our VP was applied during this phase in the first 2500 training steps, where every 500 steps we pruned 10% of the total model weight. Next, we applied the second NSP phase, also as suggested by the original paper. In this phase the model was already sparse at a level of 50%. Here to, we used the same DGX-A100 machine and trained for 1563 training steps with a batch size of 8 samples per GPU. We used a learning rate of 4e-3 and a maximum sequence length of 512.

After pre-training, we used the sparse model to fine-tune it on various of GLUE tasks as well as SQuAD1.1 via the Hugging Face Transformers library. All hyper parameter are available in the source code and will be published as a whole upon acceptance.