Applying Neural Network Compression to the Transformer

Abhi Mayur Dubal
University of Utah

UUCS-20-012

School of Computing
University of Utah
Salt Lake City, UT 84112 USA

28 August 2020

Abstract

The Transformer is a popular deep neural network model specialized for natural language processing. Like many deep neural networks, the Transformer is composed of hundreds of millions of parameters that make it favorable to undergo neural network compression techniques. Recent research has shown success with using quantization-aware training as a compression strategy for the Transformer and have delved into understanding which layers are sensitive to quantization. Moreover, existing research has used other compression strategies such as pruning but has failed to explain proper parameter tuning and the effects of these strategies on a per layer basis for the Transformer. This thesis aims to provide an in-depth analysis after applying post-training quantization, automated gradual pruning, and quantization-aware training and understanding their effects on the Transformer in hopes of improving uncompressed model accuracy while achieving high compression rates for the task of machine translation.