Quantizing the Transformer Model

Andrew Tierno

atierno@stanford.edu

Introduction

Motivation

The Transformer has become an increasingly popular choice of neural architecture for language based tasks. Perhaps most famously, Devlin et al.'s BERT contextual word embeddings have yielded state of the art results on a range of language tasks, and are constructed with a network that uses Transformers as a basic unit. **Given this rise in popularity, we aim to see if we can replicate the success of the Transformer while reducing the required space to store the model and the required time to train it through quantization of its weights and activations**

Task

To benchmark the performance of the quantized Transformer, we perform experiments on the IWSLT 2015 English-Vietnamese Dataset.

This dataset consists of segments from various TED and TEDx talks translated in both English and Vietnamese.

Approach

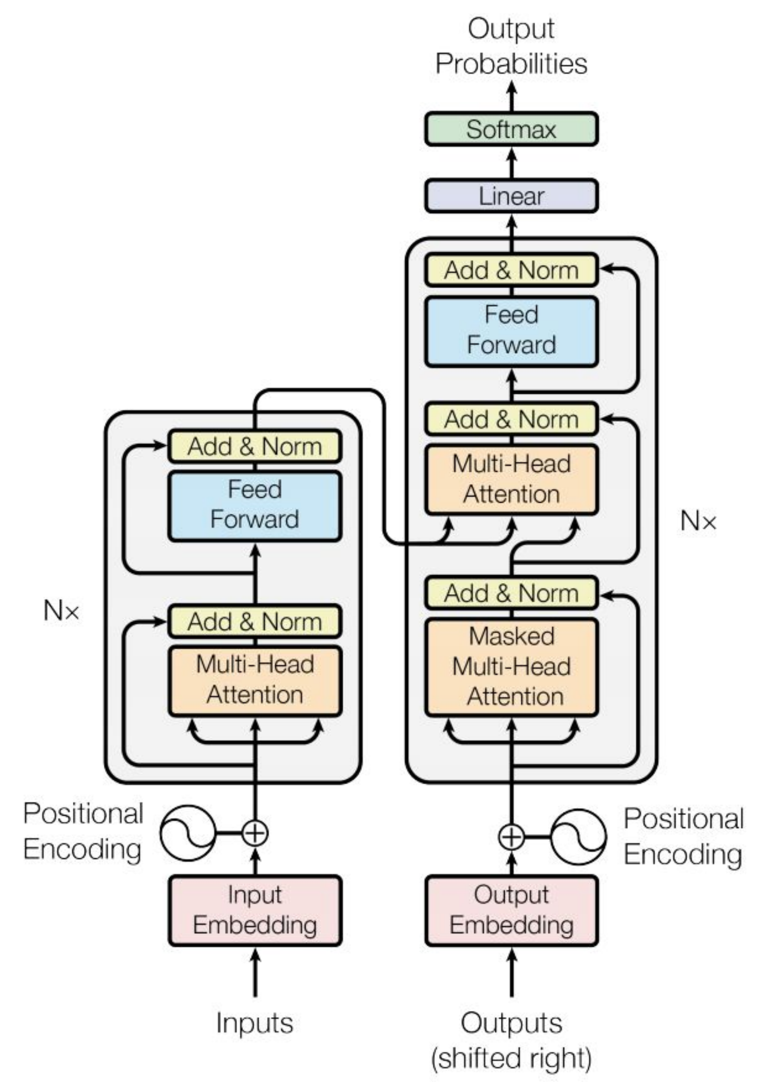
$$\text{LinearQuant}(x, bitwidth) = \text{Clip}\left(\text{round}\left(\frac{x}{bitwidth}\right)\right) \times bitwidth, minVal, maxVal)$$

$$\text{FixedQuant}(x, precision, bitwidth) = \text{Clip}(\lfloor x \times 2^{precision} \rfloor \times 2^{-precision}, minVal, maxVal)$$

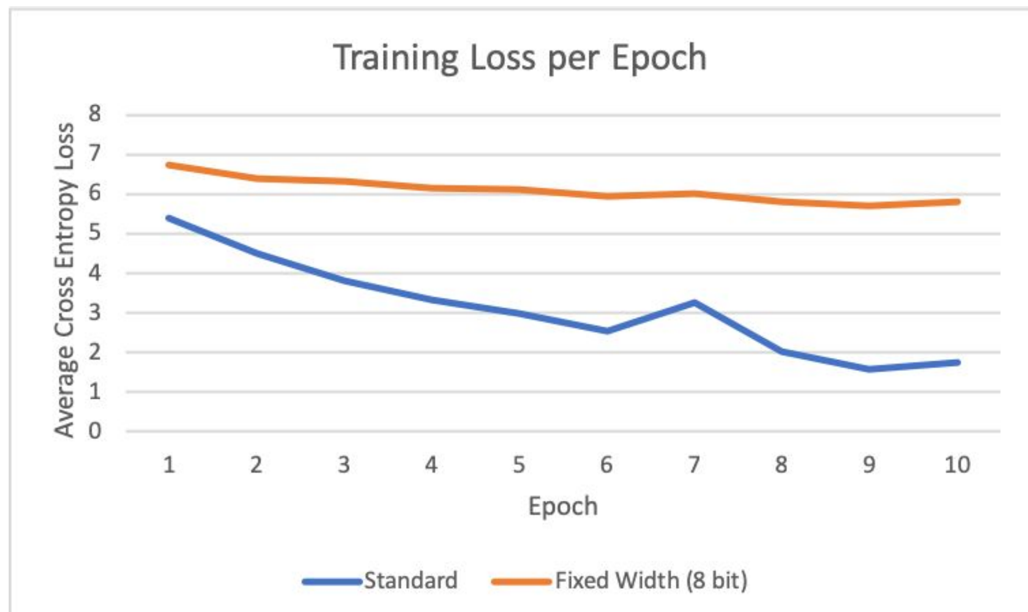
There are two primary quantization schemes that we attempt to use in this paper: linear quantization per Hubara et al. and fixed point quantization (i.e. rounding to specific precision).

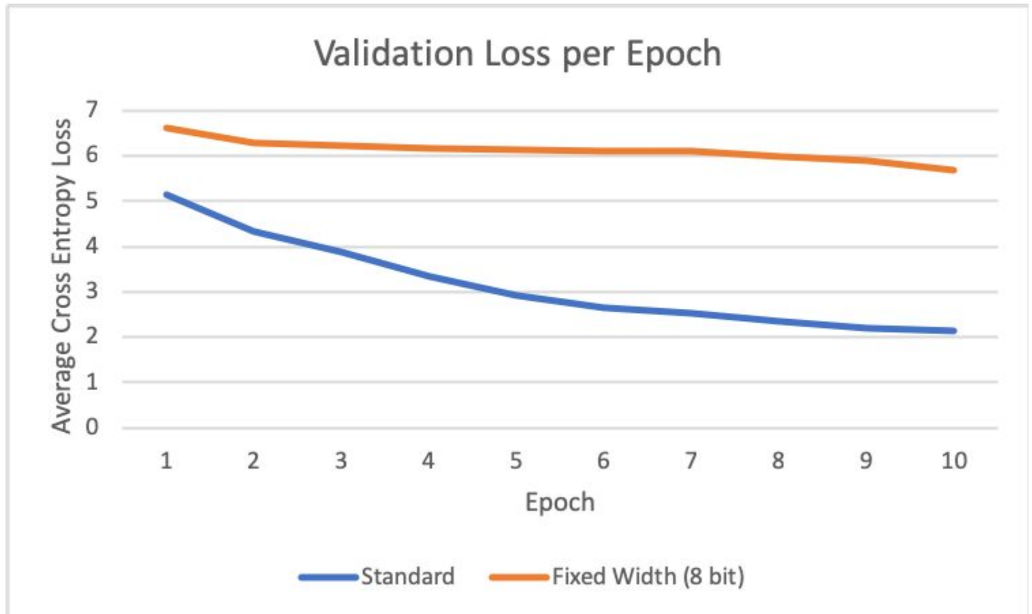
The weights of the network are stored internally in full floating point precision, but cast to their quantized forms at evaluation time. Gradients are passed straight through this quantization layer in the backpropagation step.

The Transformer



Results





Breakdown of BLEU Scores

Reference	Fixed Width (8 bit)	Linear Quant (8 bit)
20.6	10.3	2.4

Table 1: BLEU Scores for the various models

Discussion

- Both of the quantized models failed to reasonably approximate the reference models
- The two quantization schemes tended to generate many <unk> characters, suggesting that the models did not generalize well.
- The Linear Quantization model's loss was an order of magnitude larger than that of the reference or the fixed width models.

References

Niki Parmar Jakob Uskoreit-Llion Jones Aidan N. Gomez Lukaz Kaiser Illia Polosukhin Ashish Vaswani, Noam Shazeer. Attention is all you need. *NIPS*, 2017.

Yoshua Bengio, Nicholas Léonard, and Aaron C. Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *CoRR*, abs/1308.3432, 2013. URL http://arxiv.org/abs/1308.3432.

Matthieu Courbariaux and Yoshua Bengio. Binarynet: Training deep neural networks with weights and activations constrained to +1 or -1. *CoRR*, abs/1602.02830, 2016. URL http://arxiv.org/abs/1602.02830.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861, 2017. URL http://arxiv.org/abs/1704.04861.

Forrest N. Iandola, Matthew W. Moskewicz, Khalid Ashraf, Song Han, William J. Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <1mb model size. *CoRR*, abs/1602.07360, 2016. URL http://arxiv.org/abs/1602.07360.

Daniel Soudry Ran El-Yaniv Yoshua Bengio Itay Hubara, Matthieu Courbariaux. Quantized neural networks: Training neural networks with low precision weights and activations. *JMLR*, 2018.

Maximilian Lam. Word2bits - quantized word vectors. *CoRR*, abs/1803.05651, 2018. URL http://arxiv.org/abs/1803.05651.

Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/W18-6319.

Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. Xnor-net: Imagenet classification using binary convolutional neural networks. *CoRR*, abs/1603.05279, 2016. URL http://arxiv.org/abs/1603.05279.