Quantized Transformer

Chaofei Fan

stfan@Stanford.edu

Stanford
Computer Science

Introduction

Transformer is a powerful model but uses a significant amount of computing and storage resources. We use quantization to reduce model size and measure quantized model's accuracy on machine translation, sentence classification, and question answering. Our results show the following:

- 8 bits quantized model performs only **slightly worse** than 32 bits model. When fine-tuning data is scarce, 8 bits model can **outperform** 32 bits one. This means transformer can be deployed to smart phones, enabling fast offline translation.
- Use pretrained 32 bits model to initialize quantize model reduces training time and improves accuracy.
- Aggressive 1 and 4 bits quantization reduce accuracy dramatically.
- However, if only model weights are quantized, 1 bit quantization shows promising results.

References

[1] https://nervanasystems.github.io/distiller/algo_quantization/index.html

[2] Vaswani, Ashish, et al. "Attention is all you need." Advances in Neural Information Processing Systems. 2017.

[3] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

Method

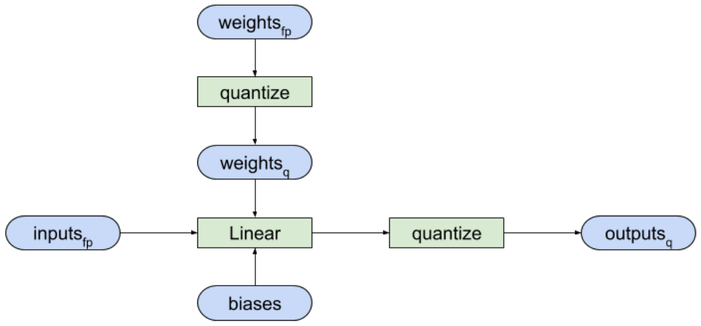


Figure 1: Linear layer with quantize operator.

Range Based Linear Quantization

$$quantize(W_{fp}, n) = round\Big((W_{fp} - min(W_{fp}))\frac{2^n - 1}{max(W_{fp}) - min(W_{fp})}\Big)$$

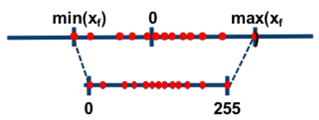


Figure 2: Range based asymmetric quantization [1]

Binary Quantization

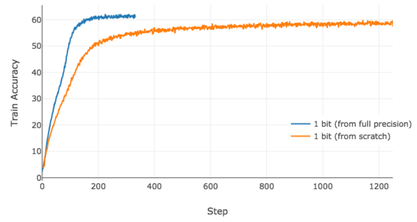
$$binarize(W_{fp}) = sign(W_{fp})$$

Straight Through Estimator

$$\frac{\partial quantize}{\partial W} = \mathbb{1} \quad \frac{\partial binarize}{\partial W} = \mathbb{1}_{|W| \leq 1}$$

Experimental Results

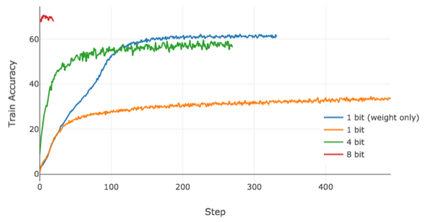
Quantize from Scratch vs. from Pretrained



WMT-14 EN-DE Machine Translation

Quantization Level	BLEU
32-bit	**27.83**
8-bit	26.94
4-bit	23.18
1-bit	2.50
1-bit (weight only)	23.88

32 bits model: Transformer base [2]



MRPC Sentence Classification

Quantization Level	Dev Acc
32-bit	84.6
8-bit	**86.3**
4-bit	68.4

32 bits model: BERT base [3]

SQuAD 1.1 Question Answering

Quantization Level	Dev EM	Dev F1
32-bit	**81.18**	**88.52**
8-bit	81.05	88.37

32 bits model: BERT base [3]