transformers.zip:
Compressing Transformers via Pruning and Quantization

Robin Cheong, Robel Daniel
{robinc20, robeld}@stanford.edu

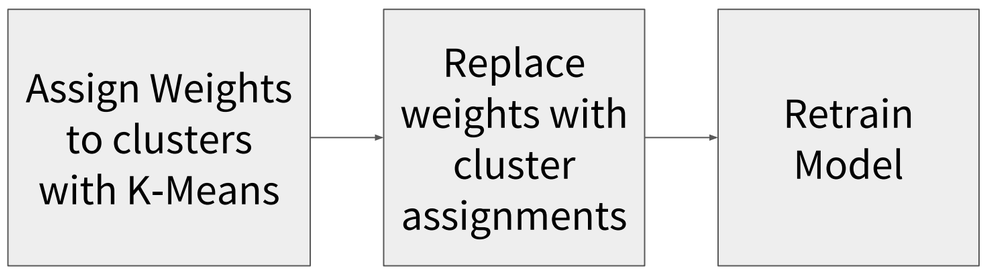
motivation

- The best pre-trained NLP models are extremely large
- This limits research to those who can afford large GPUs
- Model compression of Transformers offers a solution
- Only pruning / knowledge distillation has been tried, and very recently
- Quantization and compressed self-attention analysis have not been conducted

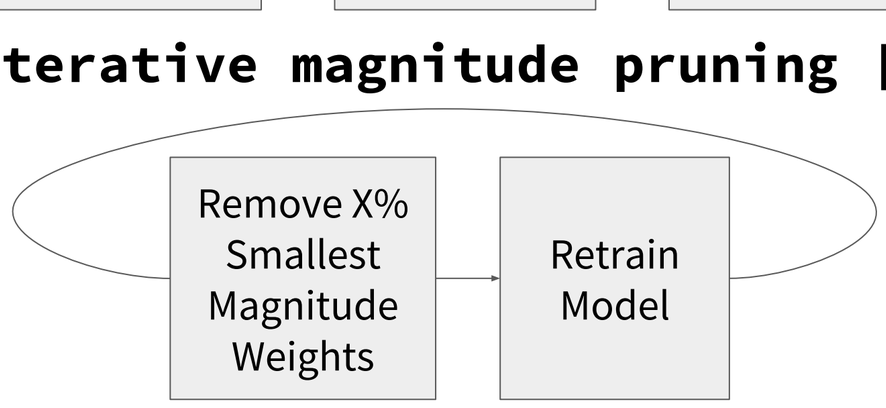
dataset

- WMT English-German Translation
- 4.5 million sentences to train on and approximately 3000 sentences to validate and test on
- BLEU score to evaluate performance/compression trade-off

k-means quantization [1]



iterative magnitude pruning [1]



flexible binary scheme (BS-Flex)

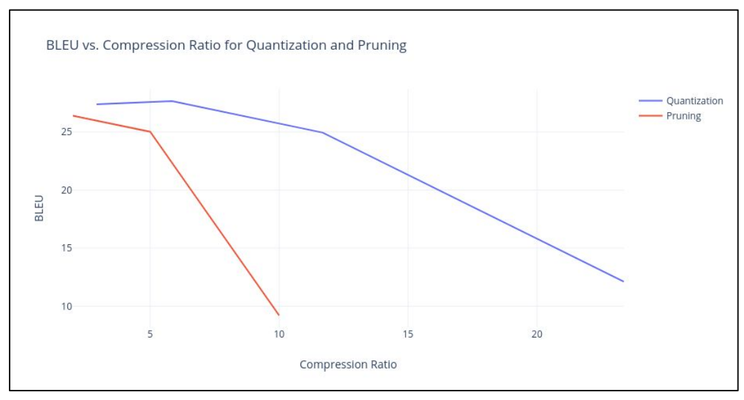
- Improve on a 1-bit quantization scheme [2]
- Allows reassignment of clusters during retraining unlike K-Means
- Set values of weights based on the average of the two centroids c_1 and c_2

$$w_{bin,ij} = \begin{cases} c_1 \text{ if } w_{ij} > \frac{c_1+c_2}{2} \\ c_2 \text{ if } w_{ij} \leq \frac{c_1+c_2}{2} \end{cases}$$

- We also experimented with a scheme that fixed the centroids, which is the original way [2].

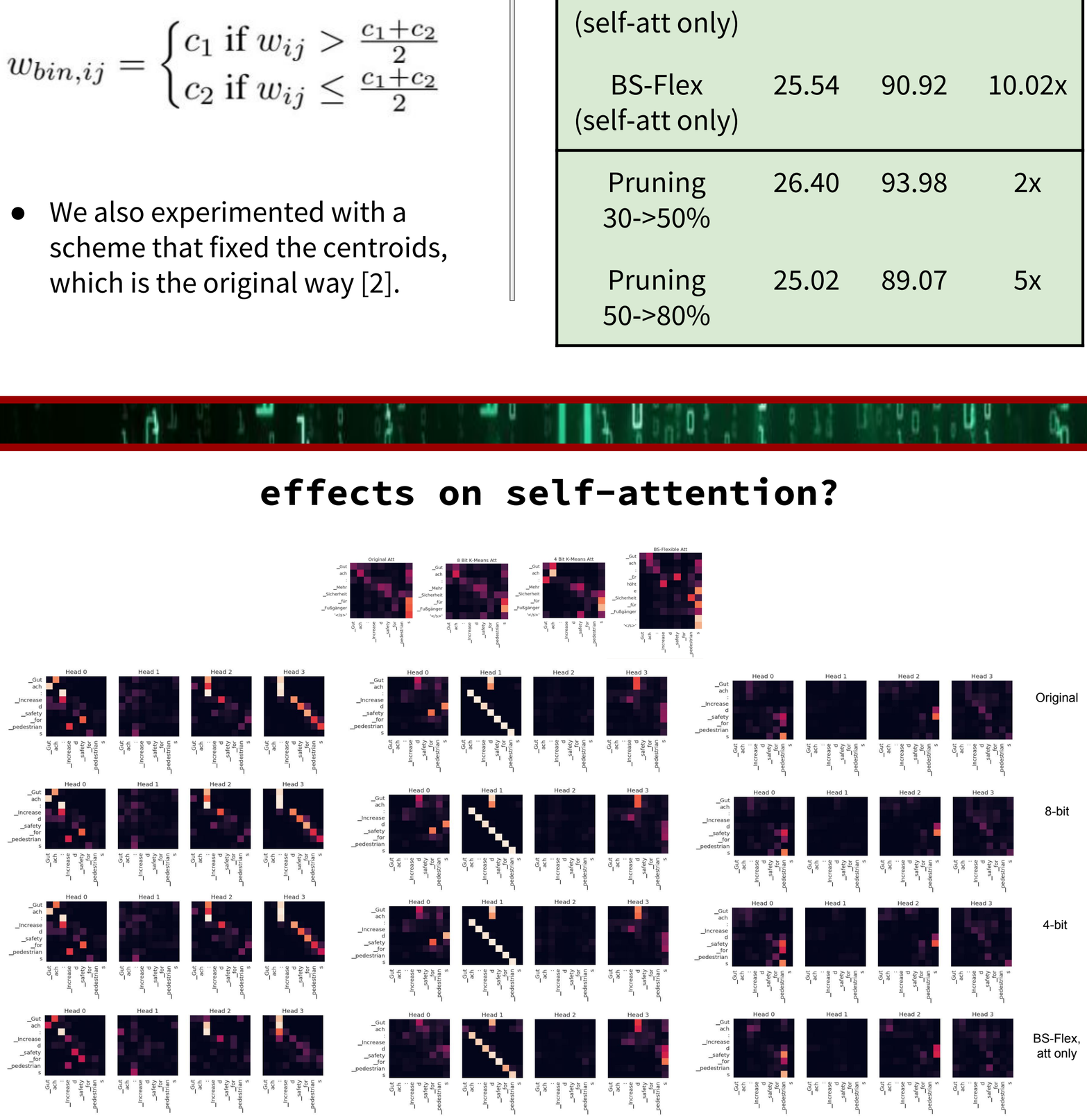
selected results

Model	BLEU	% Perf	CR
Original Transformer	28.09	100	1x
K-means (KM) 4-bit	27.65	98.43	5.85x
KM 1-bit	12.07	42.96	23.37x
KM 1-bit (self-att only)	24.96	88.85	10.02x
BS-Flex (self-att only)	25.54	90.92	10.02x
Pruning 30->50%	26.40	93.98	2x
Pruning 50->80%	25.02	89.07	5x



- Compression works really well!
- **Our binarization scheme performs as well or better than K-Means and is more than 10x faster due to PyTorch indexing**
- Pruning seems to do much worse than quantization, harder to train
- Binarizing only attention layers still gives 90% of performance!

effects on self-attention?



From top to bottom, the rows represent original, 8-bit, 4-bit, and BS-Flex (att only)

conclusion

- Significant Transformer compression can be achieved with minimal loss in performance
- This demonstrates the potential for quality compression of state-of-the-art NLP architectures
- Our binarization scheme runs faster and performs better than standard 1-bit quantization
- Self-attention is highly resistant to quantization; replacing elements with only 16 values produces a nearly identical distribution

references

[1] Song Han, Huizi Mao, and William J. Dally. “Deep Compression: Compressing Deep Neural Network with Pruning, Trained Quantization and Huffman Coding”.
[2] Maximilian Lam. “Word2Bits - Quantized Word Vectors”.
[Codebase] - Guillaume Klein et al. “OpenNMT: Open-Source Toolkit for Neural Machine Translation”.