Weight Fixing Networks (WFN)
Single Codebook Neural Network Quantisation

Christopher Subia-Waud

UK Research and Innovation

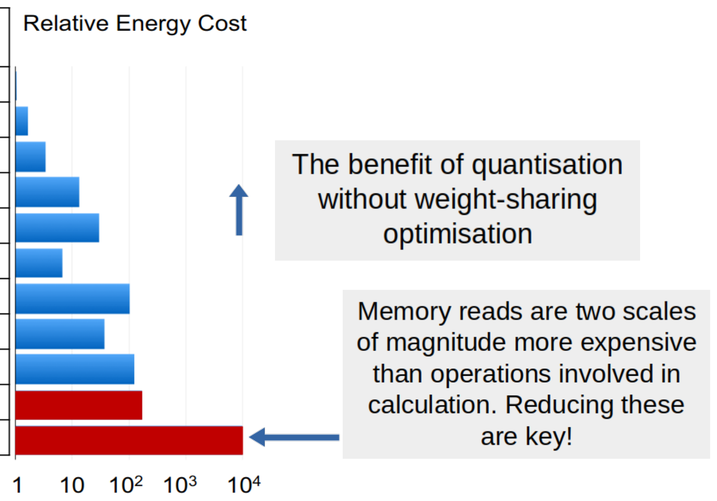
University of Southampton

The Big Picture

There is a large gap between the metrics used in evaluating model-reduction techniques and the ultimage goal of energy efficiency and reduced inference latency.

Inference op costs

Operation:	Energy (pJ)
8b Add	0.03
16b Add	0.05
32b Add	0.1
16b FP Add	0.4
32b FP Add	0.9
8b Multiply	0.2
32b Multiply	3.1
16b FP Multiply	1.1
32b FP Multiply	3.7
32b SRAM Read (8KB)	5
32b DRAM Read	640



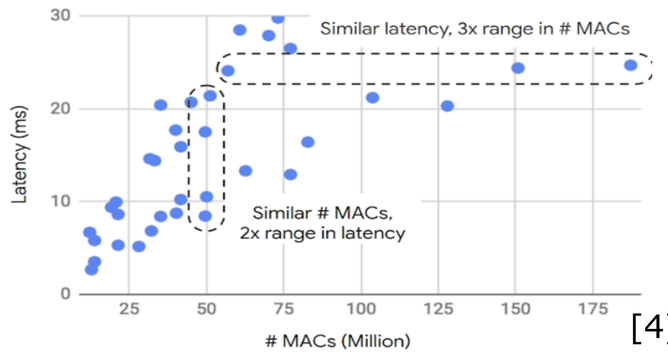
The benefit of quantisation without weight-sharing optimisation

Memory reads are two scales of magnitude more expensive than operations involved in calculation. Reducing these are key!

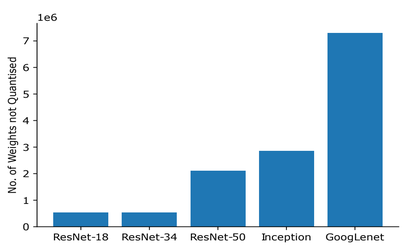
[3]

Three Observations

1. Fewer MAC ops doesn't always translate into lower latency & energy costs

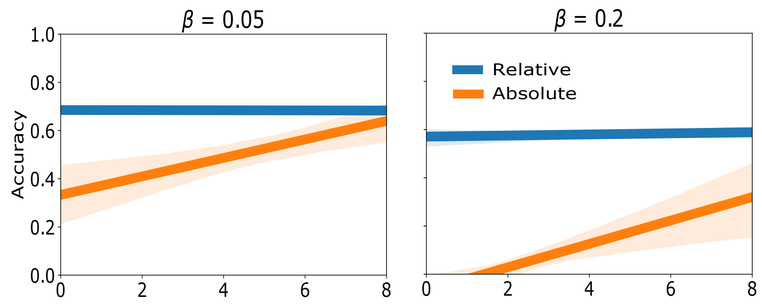


[4]



2. Quantisation works do not fully quantise lots of network and have different codebooks for each layer

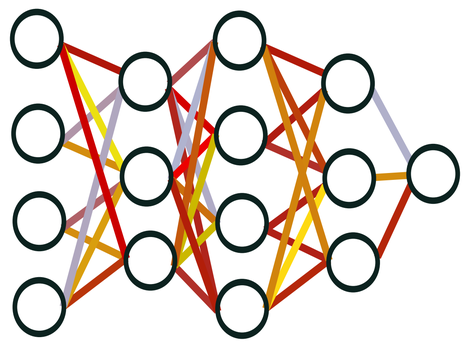
3. Networks are more resilient to relative (muliplicative) than absolute (additive) noise.



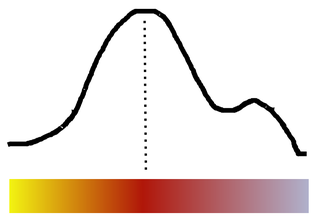
Proposed Method for Energy Efficient Inference

A focus on reducing memory reads through a reduction in the number of unique weights

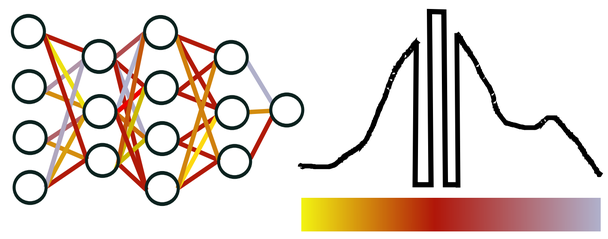
Start with a fully converged model, here we represent different weight values with a colour



Select the largest k clusters from the weight-space distribution. We Increase k as needed to allow for more weights to be clustered without moving weights too far from their converged values



Assign the weights that are close to the cluster centroids to their centroid and fix them (no more learning for these weights)



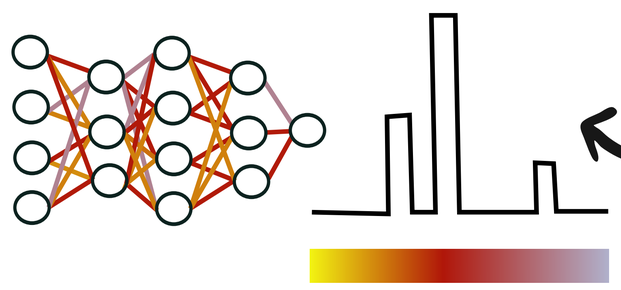
Re-train the network for a few epochs, allowing the un-clustered weights to move. We additionally add a regularisation term to encourage un-clustered weights to be close to the cluster centroids. Repeat these first three steps a few times.

No of Un-clustered weights No of cluster centers

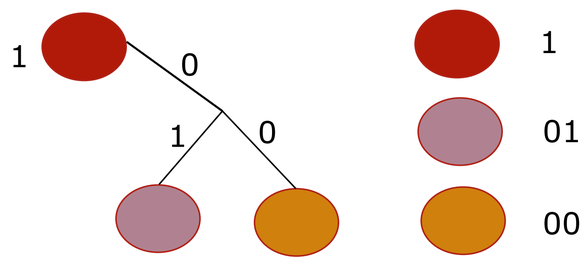
$$\gamma \sum_{i \in W_{free}}^{N} \sum_{j}^{k} D^{+}_{reg}(w_i, c_j) p(c_j|w_i)$$

Relative Quantisation Distance

The procedure is complete once all weights have been fixed to one of k cluster centers



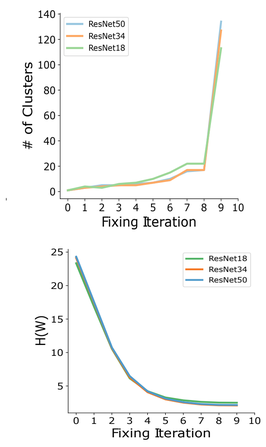
Huffman encoding can be used to further compress the bit-space required to represent the network weight indices



Results

Here we present an initial experiment using WFN to compress ResNet networks trained on the Imagenet dataset and compare with a SOTA quantisation approach, APoT.

Model	Method	Accuracy (%) Top-1	Top-5	Full Network Entropy	Full Network Param Count	No BN Entropy	No BN Param Count	No BN-FL Entropy	No BN-FL Param Count	Model Size	CR
ResNet-18	Baseline	68.9	88.9	23.3	10756029	23.3	10748288	23.3	10276369	46.8MB	1
	APoT (3bit)	69.9	89.2	5.77	9237	5.76	1430	5.47	274	4.56MB	10.2
	WFN ($\gamma = 0.015$)	67.3	87.6	2.72	90	2.71	81	2.5	81	3.5MB	13.37
	WFN ($\gamma = 0.01$)	69.7	89.2	3.0	164	3.0	153	2.75	142	3.8MB	12.3
	WFN ($\gamma = 0.0075$)	70.3	89.1	4.15	193	4.13	176	3.98	162	4.6MB	10.2
ResNet-34	Baseline	73.3	91.2	24.1	19014310	24.1	18999320	24.1	18551634	87.4MB	1
	APoT (3bit)	73.4	91.1	6.77	16748	6.75	16474	6.62	389	8.23MB	10.6
	WFN ($\gamma = 0.015$)	72.2	90.9	2.83	117	2.81	100	2.68	100	6.9MB	12.6
	WFN ($\gamma = 0.01$)	72.6	91.0	3.48	164	3.47	132	3.35	130	7.9MB	11.1
	WFN ($\gamma = 0.0075$)	73.0	91.2	3.87	233	3.85	191	3.74	187	8.5MB	10.3



References

[1] Yang, Tien-Ju, Yu-Hsin Chen, and Vivienne Sze. "Designing energy-efficient convolutional neural networks using energy-aware pruning." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[2] Iandola, Forrest N., et al. "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size." arXiv preprint arXiv:1602.07360 (2016).

[3] Horowitz, Mark. "1.1 computing's energy problem (and what we can do about it)." 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC). IEEE, 2014.

[4] https://ai.googleblog.com/2018/04/introducing-cvpr-2018-on-device-visual.html

MINDS·CDT