Quantization for Energy Efficient Convolutional Neural Networks

Joao Vitor Mascarenhas[1], Chao Yang (Advisor) [2], Daniela Ushizima (Advisor) [2]

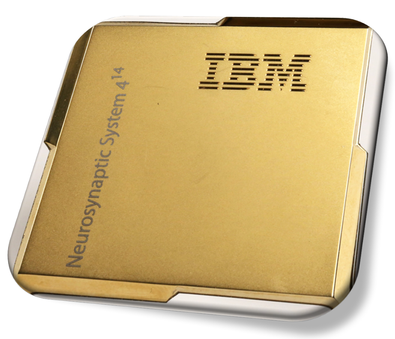
[1]Federal University of Ouro Preto, [2]Lawrence Berkeley National Laboratory

ABSTRACT

A traditional Convolutional Neural Network (CNN) is parameterized by floating point weights and biases and takes floating point data as input. In many cases, the floating point representation of these parameters and input is more than necessary. The use of a more compact representation of the parameters and input allows CNNs to be deployed on energy efficient architectures that operate with a few bits and much lower memory footprint. This work focuses on data reduction and quantization schemes that can be applied to a trained CNN for classifying scientific simulation data. We show that each neuron and synapse can be encoded with only one byte to maintain accuracy above 98%.

CONVOLUTIONAL NEURAL NETWORKS

- Most CNNs are built to traditional computing platforms.
- The training algorithm is typically carried out on the floating point arithmetic unit of CPUs and GPUs.
- Recent studies indicate that most CNN parameters are redundant.
- It is possible to lower the precision of the CNN parameters and input data without significantly changing success rate of the CNN classification.
- This allows the use of fewer bits to represent parameters and input data from scientific experiments produced through Spider simulation.



METHODS

Quantization strategy using K-means clustering algorithm with Lloyd's optimization:

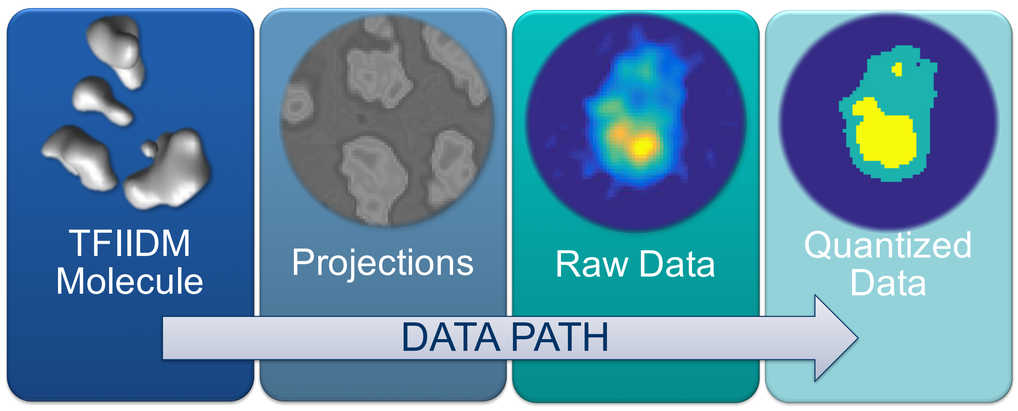
- Quantizing single-precision floating point input data, CNN weights, and biases to a much smaller number of levels that can be represented by a fewer bits and a code book (or lookup table.)
- Reduction from **single precision** representation **to** at most **one byte**.

Train-then-constrain approach:

- Train a CNN with the input data quantized.
- Modify the trained CNN by quantizing the weights and biases.

EXPERIMENTS AND RESULTS

- To test the effect of quantization, we used a dataset with 2,500 simulated cryo-electron microscopy (CryoEM) images of the TFIID molecule [1].
- The quantized CNN is used to classify CryoEM images into 84 different classes based on the orientation of the projection.
- The used CNN consists of 6 layers, constructed with MatConvNet.
- We used 80% of the data for training and the 20% were used to test the correctness of CNN classification.



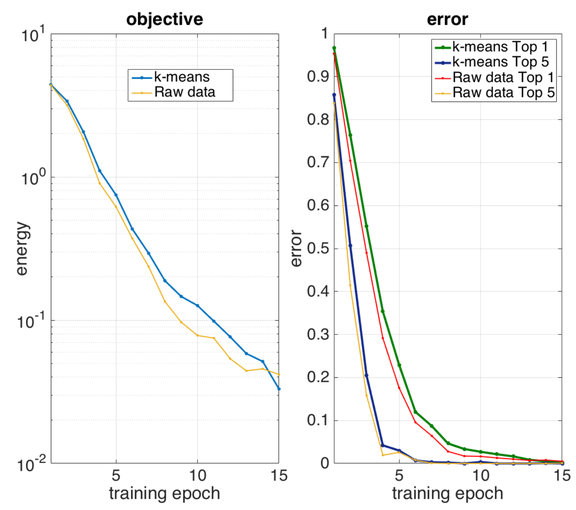
- Quantizing both the data and the weights into 3-levels, **the success rate is 92%**.
- As expected, the success rate improves as the number of quantization levels increase.
- Without quantization, the success rate of the classification is 100%.

CNN success rate (%) for different quantization levels

Weights quantization levels

Data quantization levels	1	2	3	4	5	6	7	8
NA	2.2	71.2	84	90.6	96.4	98	97.8	98.4
1	2.2	1.4	1.6	1	1	1	1.4	1
2	2.2	78.6	90.2	94.8	95.6	96.2	97	96.8
3	2.2	82.4	**92.4**	97	94.8	97.8	97.6	98
4	2.2	65.4	88.2	90.6	96	97.6	97	98.8
5	2.2	73	87	92.6	95.6	95.6	97.4	97.6
6	2.2	62	84.2	92.6	93.4	94.8	95.2	**99**
7	2.2	67	81	91.8	93.4	94.6	96.8	98.2
8	2.2	71.6	86.2	88.8	96	96.4	97.8	98.8

TRAINING THE CNN



- During training, the error and objective curves for the quantized data had very similar accuracy to the original (not quantized) data.

IMPACT OF LOW PRECISION REPRESENTATION

Such reduction in the number of bits allows transferring computation to energy efficient processors for deep learning. Devices such as the IBM's TrueNorth chip and Google's tensor processing unit (TPU) strive computationally under such constraints. In addition, allows:

- faster data classification;
- data and model compression;
- data smoothing, circumventing noise artifacts.

REFERENCES

[1] F. Andel, A. G. Ladurner, C. Inouye, R. Tjian, and E. Nogales. Three-dimensional structure of the human tfiid-iia-iib complex. Science, 286(5447):2153–2156, 1999.

[2] S. K. Esser, P. A. Merolla, J. V. Arthur, A. S. Cassidy, R. Appuswamy, A. Andreopoulos, D. J. Berg, J. L. McKinstry, T. Melano, D. R. Barch, C. di Nolfo, P. Datta, A. Amir, B. Taba, M. D. Flickner, and D. S. Modha. Convolutional networks for fast, energy-efficient neuromorphic computing. CoRR, abs/1603.08270, 2016.

[3] J. Frank. Three-Dimensional Electron Microscopy of Macromolecular Assemblies: Visualization of Biological Molecules in Their Native State. Oxford University Press, 2006.

[4] Y. Gong, L. Liu, M. Yang, and L. D. Bourdev. Compressing deep convolutional networks using vector quantization. CoRR, abs/1412.6115, 2014.

[5] S. Gupta, A. Agrawal, K. Gopalakrishnan, and P. Narayanan. Deep learning with limited numerical precision. CoRR, abs/1502.02551, 2015.

[6] S. Lloyd. Least squares quantization in pcm. IEEE Trans. Inf. Theor., 28(2):129–137, Sept. 2006.

[7] M. Tu, V. Berisha, Y. Cao, and J. Seo. Reducing the model order of deep neural networks using information theory. CoRR, abs/1605.04859, 2016.

PROCESS OVERVIEW