Minimum Precision Requirements for Deep Learning with Biomedical Datasets

Charbel Sakr
University of Illinois at Urbana-Champaign
sakr2@illinois.edu

Naresh Shanbhag
University of Illinois at Urbana-Champaign
shanbhag@illinois.edu

***Abstract*—Deep neural networks (DNNs) are powerful machine learning models but are typically deployed in large computing clusters due to their high computational and parameter complexity. Many biomedical applications require embedded inference on resource-constrained platforms thus causing a challenge when considering the deployment of DNNs. One method to address this challenge is via reduced precision implementations. We use an analytical method to determine suitable minimum precision requirements of DNNs and show its application to the CHB-MIT EEG seizure detection dataset and the Bonn dataset for brain electrical activity recognition. We show that our method leads to $2\times$ reduction in average precision and 45% complexity reduction compared to the minimum uniform precision assignment. Compared to a conventional 16-b precision assignment, our method leads to $9\times$ complexity reduction. Furthermore, we study the impact of network topology on precision and accuracy. Once again we find our method to be $\sim 2\times$ more efficient that the uniform assignment for all topologies considered.**

***Index Terms*—neural networks, deep learning, precision, accuracy, complexity**

I. Introduction

Neural networks and deep learning are achieving considerable accuracy in many inferential tasks. The most popular applications of deep learning are cognitive such as computer vision [1], speech [2], and language [3]. However, the strong representational power of neural networks also makes them suitable for biomedical-based machine learning tasks. Indeed, raw biological data is being collected in large volumes today [4] soliciting the need for *information/knowledge extraction*. Consequently, deep learning is currently being employed for several biomedical applications such as medical imaging [5], biological anomaly and pathology detection [6], and early diagnosis of diseases [7], to name a few.

However, due to their biomedical nature, many of the above applications require inference to be done in-situ, in an embedded platform close to the patient. Unfortunately, the high complexity of neural networks makes them hard to deploy onto such *resource-constrained* platforms. Thus, most works on embedded biomedical circuits and systems have so far focused on the mapping of classical algorithms. For instance, a low-power digital implementation of the Pan Tompkins Algorithm was implemented for heart rate monitoring [8]. Similarly, a low complexity architecture for seizure detection was designed based on incremental precision FFT-based feature extraction [9]. Another popular algorithm is the support vector machine which was used in a micro power complete system-on-a-chip (SoC) design for EEG-based seizure detection [10].

This work was supported in part by C-BRIC, one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA.

Therefore, a gap exists between state-of-the-art machine learning models such as deep neural networks (DNNs) and their mapping onto embedded biomedical platforms. Closing this gap necessitates a fresh look at complexity reduction techniques for DNNs. One fundamental reason contributing to the high complexity of DNNs is their implementation in 32-b floating-point arithmetic on CPUs and GPUs. Thus, reduced precision representations, such as 16-b fixed-point [11], have been employed to reduce the cost of implementation of DNNs. It was shown to be possible to directly train 16-b fixed-point networks [12]. This approach necessitates a discrete optimization and has no convergence guarantees. Moons et al. [13] empirically determine the optimal precision by training an ensemble of networks with varying topologies and datasets. They find that typically decreasing the precision from 16-b or 8-b down to 4-b has little effect on accuracy but leads to great complexity reduction. Accuracy degradation is observed when precision is decreased to 2-b or 1-b.

The inherent robustness of DNNs suggests that quantizing a pre-trained model leads to a much easier, yet reliable way of reducing the complexity. When considering quantization of pre-trained networks, the main challenge is to determine a suitable precision configuration. Furthermore, due to the very large design space, it is prohibitive to perform an exhaustive search [14]. Moons et al. [15] perform a grid search where quantization is applied one layer at a time. The corresponding precision is found to be dependent on the order of the search because the noise budget is exhausted at each iteration. Thus, dependence on data and network statistics is not captured. Therefore precision assignment needs to be done *analytically*. The solution in [16] is one where precision assignment is chosen so that to satisfy a signal-to-quantization-noise ratio (SQNR) condition at the output. This solution suffers from SQNR not being a meaningful metric in the context of classification and machine learning. In [17], a theory was developed whereby it was possible to analytically quantify the accuracy degradation due to network quantization as a function of precision. This theory was used to obtain an efficient precision assignment methodology [18].

In this work, we employ the aforementioned analysis to efficiently quantize pre-trained networks on the CHB-MIT EEG dataset for seizure detection [19] and the Bonn dataset for brain electrical activity recognition (BEAR) [20]. Unlike image-based datasets such as ImageNet [1], we find activation precision requirements to be dominant. We also compare the results of our proposed method to the minimum uniform precision assignment and show up to $\mathbf{2}\times$ average precision

reduction and **45%** complexity reduction. Compared to a conventional 16-b precision assignment, our method leads to $\mathbf{9}\times$ complexity reduction. We further investigate the trade-offs between network topology, precision, and complexity and find that the benefits of our method generalize across network topologies.

The rest of this paper is organized as follows: Section II provides background on our proposed method and complexity metrics. Section III includes applications of the proposed method to biomedical datasets as well as accuracy vs. precision vs. complexity evaluations for fixed and varying network topologies. Finally, we conclude our paper in Section IV.

II. Background

A. Precision Analysis of Deep Neural Networks

We consider a general feedforward neural network with L layers. Let $\{\mathcal{A}_l\}_{l=1}^{L}$ and $\{\mathcal{W}_l\}_{l=1}^{L}$ be the layer-wise partitions of activations and weights, respectively. We assume all activations and weights are quantized to fixed-point, with per-layer precisions $\{B_{A,l}\}_{l=1}^{L}$ and $\{B_{W,l}\}_{l=1}^{L}$, respectively. Furthermore, the accuracy metric we utilize is the *mismatch probability*: $p_m = P\left(\hat{Y}_{fx} \neq \hat{Y}_{fl}\right)$, which is the probability that the predicted label $\hat{Y}_{fx}$ of a fixed-point neural network is different from $\hat{Y}_{fl}$, that of its floating-point counterpart. It can be shown that [18]:

$$p_m \leq \sum_{l=1}^{L} \left(\Delta_{A,l}^2 E_{A,l} + \Delta_{W,l}^2 E_{W,l}\right) \tag{1}$$

where $\Delta_{A,l} = 2^{-(B_{A,l}-1)}$ and $\Delta_{W,l} = 2^{-(B_{W,l}-1)}$ are the activation and weight quantization step-sizes at layer l, respectively, and

$$E_{A,l} = \mathbb{E}\left[\sum_{\substack{i=1 \\ i\neq\hat{Y}_{fl}}}^{M} \frac{\sum\limits_{a\in\mathcal{A}_l}\left|\frac{\partial(Z_i - Z_{\hat{Y}_{fl}})}{\partial a}\right|^2}{24|Z_i - Z_{\hat{Y}_{fl}}|^2}\right]$$

and

$$E_{W,l} = \mathbb{E}\left[\sum_{\substack{i=1 \\ i\neq\hat{Y}_{fl}}}^{M} \frac{\sum\limits_{w\in\mathcal{W}_l}\left|\frac{\partial(Z_i - Z_{\hat{Y}_{fl}})}{\partial w}\right|^2}{24|Z_i - Z_{\hat{Y}_{fl}}|^2}\right]$$

are the activation and weight quantization noise gains at layer l, respectively. Note the dependence of the quantization noise gains on the number of classes (M) and the soft outputs ($\{Z_i\}_{i=1}^{M}$).

Interestingly, the noise gains can be obtained as part of a standard back-propagation procedure and need to be computed only once making (1) very practical. Furthermore, observe that (1) is a sum of $2L$ terms. The design parameters are the $2L$ precision assignments, $\{B_{A,l}\}_{l=1}^{L}$ and $\{B_{W,l}\}_{l=1}^{L}$. In order to minimize the precision, the sum has to first be balanced. To do so, the minimum quantization noise gain is first computed:

$$E_{\min} = \min\left(\{E_{A,l}\}_{l=1}^{L}, \{E_{W,l}\}_{l=1}^{L}\right). \tag{2}$$

Then, a reference minimum precision $B_{\min}$ is chosen, and for each layer l, the precision is set as follows:

$$B_{A,l} = round\left(\log_2\left(\sqrt{\frac{E_{A,l}}{E_{\min}}}\right)\right) + B_{\min} \tag{3}$$

and

$$B_{W,l} = round\left(\log_2\left(\sqrt{\frac{E_{W,l}}{E_{\min}}}\right)\right) + B_{\min} \tag{4}$$

Note that at least one of the $2L$ precision assignments will equal $B_{\min}$. Observe that (1)-(4) can be used to alleviate the need for trial-and-error, and to reduce the search space from a $2L$ dimensional grid to just a one dimensional axis.

B. Complexity Metrics

In order to quantify the benefits of the proposed precision reduction method, we shall consider two measures of complexity [17]: *computational* and *representational* costs.

The computational cost is the total number of full adders (FAs) needed per decision and is given by:

$$\sum_{l=1}^{L}\Big[N_l\left(D_l\, B_{A,l} B_{W,l} + (D_l - 1)(B_{A,l} + B_{W,l} + \lceil\log_2(D_l)\rceil - 1)\right)\Big]$$

where N_l and D_l are the number and dimensionality of dot products computed at layer l, respectively.

The representational cost is the total number of bits needed to represent both weights and activations, and is given by:

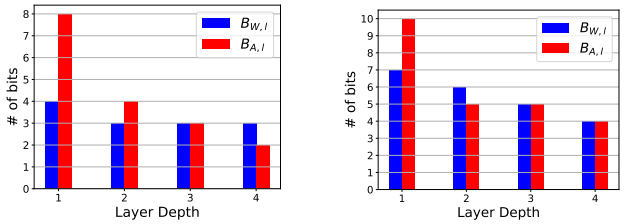
$$\sum_{l=1}^{L}\left(|\mathcal{A}_l|\, B_{A,l} + |\mathcal{W}_l|\, B_{W,l}\right)$$

III. Precision Analysis with Biomedical Datasets

A. Datasets

We employ the above precision analysis on two datasets: the **CHB-MIT** EEG dataset for seizure detection [19] and the Bonn dataset for Brain Electrical Activity Recognition (**BEAR**) [20].

The CHB-MIT dataset consists of EEG recordings from pediatric subjects with intractable seizures. The dataset was built by monitoring 24 patients for different durations. Each of the recordings is obtained by reading data from 23 electrical channels. Along with the recordings are provided the times where seizures occurred. To make the dataset compatible with the usual machine learning setup, we sample feature vectors by collecting the readings from 23 channels for 20 consecutive time intervals. Hence, the feature vectors have a dimension of 460. The corresponding labels are naturally binary and represent the event, or lack thereof, of a seizure. Finally, the dataset used is obtained by sampling 5000 random elements from the resulting data points. It is to be noted that seizures are extremely rare events (happening less than 1% of the time), thus, to decrease the bias of the dataset, the samples are chosen such that approximately 10% correspond to seizures. The dataset is split into training and testing sets by using a random 80%/20% partition.



(a) (b)

Fig. 1: Representation of the minimum per-layer precision configuration such that $p_m \leq 1\%$ for the DNNs on the (a) CHB-MIT and (b) BEAR datasets.

The BEAR dataset can be obtained from the UCI Machine Learning repository [21]. This dataset is also based on electrical readings of patients, but was pre-processed and re-structured by its authors. Each data sample consists of a 178 dimensional feature vector and a label corresponding to one of five classes: (1) eyes closed, (2) eyes open, (3) no seizure activity but detection of tumor, (4) no seizure activity and tumor detected in opposite brain hemisphere, and (5) seizure activity. Due to the larger number of classes, the associated classification task is harder. Overall, the dataset contains 11500 samples and is once again split into a random 80%/20% for training and testing.

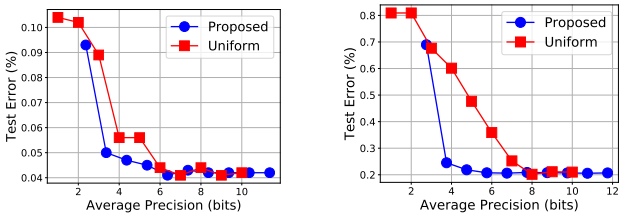
B. Neural Networks and Precision Assignments Considered

For both datasets, we consider the same DNN topology which we describe as: $F - W - W - W - M$, where F is the dimension of the input feature vector, W is the *internal network width* (or number of hidden activations per layer), and M is the number of soft outputs (which equals the number of classes). The chosen value of W is 400 unless otherwise specified. The activation function used is a ReLU-like clipping activation function with a saturating upper level of 2. Each network is trained for 1000 epochs using momentum-SGD to optimize a cross-entropy loss function with an initial learning rate of 0.1 that is decreased by a factor of 10 every 300 epochs. The only regularization applied is a 10% dropout on the hidden activations. The converged floating-point accuracy is of **4.2%** and **20.70%** test error for the CHB-MIT and BEAR datasets, respectively. These two accuracies obtained via floating-point training are of fine quality for the corresponding datasets and are used as reference for the remainder of the paper. For the above pre-trained networks, we consider three types of precision assignments:

Proposed precision assignment: minimum per-layer precision configuration using the methodology described by (1)-(4) such that $p_m \leq 1\%$. This precision assignment is found via a search on $B_{\min}$ in (3) & (4).

Uniform precision assignment: minimum identical precision for all activations and weights such that $p_m \leq 1\%$. This precision assignment is found via a search.

Conventional precision assignment: setting all activation and weight precisions to 16-b as is typically done in digital DNN implementations [11].



(a) (b)

Fig. 2: Test error vs. average precision when using the proposed precision analysis and comparison with uniform precision assignment for the DNNs on the (a) CHB-MIT and (b) BEAR datasets.

TABLE I: Test error, computational and representational costs vs. precision configuration using proposed, uniform, and conventional precision assignments for the DNNs on the CHB-MIT and BEAR datasets.

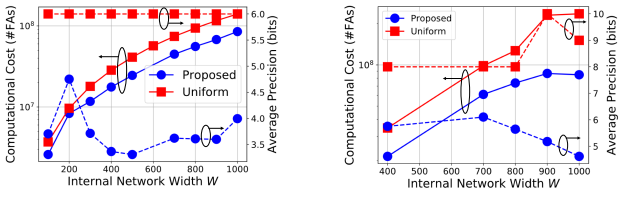
Precision Configuration	Test Error	Computational Cost (MFAs)	Representational Cost (Mbits)
CHB-MIT			
Proposed (Fig. 1(a))	5.0%	17.6	1.72
Uniform (6-b)	4.4%	28.2	3.06
Conventional (16-b)	4.2%	150	8.15
BEAR			
Proposed (Fig. 1(b))	20.74%	21.5	2.28
Uniform (8-b)	20.18%	34.7	3.17
Conventional (16-b)	20.69%	117	6.35

C. Precision vs. Accuracy vs. Complexity

Figure 1 shows a representation of the precision configuration resulting from our proposed assignment for both CHB-MIT and BEAR cases. It has been shown [18] that for image-based datasets, the precision requirements of weights are higher than those of activations. This is not the case for biomedical data as shown in Fig. 1: *activation precision requirements dominate*. This result is unsurprising considering that image data is usually quite redundant and thus robust to input quantization noise. One the other hand, quantization of biomedical signals filters out important information. Additionally, it can be seen *that precision requirements decrease as layer depth increases*, a trend that was also observed for image-based datasets. Moons et al. [15] perform a grid search whereby layers are quantized successively. Maximal precision reduction depends on the order of the search and is achieved for the first layer quantized, due to the exhaustion of the noise budget of the network. Our method, being analytical, determines all precisions simultaneously, causing the inter-layer precision trends to be dependent on the data and network statistics.

In Fig. 2, we show how accuracy varies as a function of *average precision*, which is the average precision of all weights and activations. The plot is obtained by sweeping the value of $B_{\min}$ in (3)-(4) and recording test error and average precision for each case. For comparison, we also plot test error vs. precision when using a uniform precision assignment. The proposed method is clearly superior to the latter and leads to a $\mathbf{2}\times$ (6/3 and 8/4) reduction of average precision while maintaining a similar level of accuracy.

To quantify the benefits, computational and representational



(a) (b)

Fig. 3: Computational cost and average precision vs. internal network width for minimum precision networks achieving a test error less than 5% on the (a) CHB-MIT dataset and 21% on the (b) BEAR dataset.

costs are reported along with test error in Table I. While the accuracy levels are similar, the proposed approach reduces the complexity by $\sim \mathbf{45}\%$ compared to the uniform assignment. There is a consistent reduction in computational and representational costs by 10 MFAs and 1 Mbits, respectively. Compared to the conventional case, savings are as high as $\sim \mathbf{9}\times$.

D. Complexity and Topology

The above focuses solely on reducing the cost of implementation of a pre-trained model with *fixed* topology. In order to better understand the effects of topology on complexity and precision, we run an experiment whereby we repeat all of the above experiments for networks with varying internal network width W. Specifically, we sweep W from 100 to 1000 in increments of 100. We quantize each resulting network using our method as well as uniformly, and for each case we retain the configurations achieving a test error less than 5% and 21% for the CHB-MIT and BEAR datasets, respectively.

For every resulting configuration, we compute computational cost and average precision and plot them against internal network width. These plots are included in Fig. 3. First, we observe is that *the benefits of our proposed method generalize across network topologies*. Indeed, our proposed method always leads to the least precision and complexity (measure in computational cost) requirements. Similar trends were observed for the representational cost but are not included due to space limitations. The average precision is once again almost $\sim \mathbf{2}\times$ lower when using the proposed method as compared to the uniform assignment.

For the proposed method, an interesting observation is the general trend of increase in complexity but decrease in average precision when internal network width increases. This suggests that when the model complexity is increased, the performance of the network is boosted making it more robust to quantization. However, in spite of the resulting precision reduction, the complexity still increases as it scales quadratically with network width. Furthermore, we observe that across datasets and network topologies, the minimum precision is typically within 4-b to 5-b, which is consistent with the empirically derived conclusion in [13], in spite of the difference in data statistics between image-based and biomedical-based datasets.

IV. Conclusion

In this paper, we have presented an analytical method to determine suitable minimum precision requirements of DNNs and showed its application to the CHB-MIT EEG seizure detection dataset and the Bonn dataset for brain electrical activity recognition. We showed that our method leads to almost half precision and complexity requirements compared to the minimum uniform precision assignment for both fixed and varying network topologies. We also showed up to $9\times$ complexity reduction comared to the conventional precision assignment.

References

[1] J. Deng *et al.*, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 248–255.

[2] D. Amodei *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International Conference on Machine Learning*, 2016, pp. 173–182.

[3] D. Bahdanau *et al.*, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[4] S. Min *et al.*, "Deep learning in bioinformatics," *Briefings in bioinformatics*, vol. 18, no. 5, pp. 851–869, 2017.

[5] H. Greenspan *et al.*, "Deep learning in medical imaging: Overview and future promise of an exciting new technique," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1153–1159, 2016.

[6] Y. Bar *et al.*, "Chest pathology detection using deep learning with non-medical training," in *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*. IEEE, 2015, pp. 294–297.

[7] S. Albarqouni *et al.*, "Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1313–1321, 2016.

[8] R. A. Abdallah and N. R. Shanbhag, "An energy-efficient ecg processor in 45-nm cmos using statistical error compensation," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 11, pp. 2882–2893, 2013.

[9] S. Koteshwara and K. K. Parhi, "Incremental-precision based feature computation and multi-level classification for low-energy internet-of-things," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 2018.

[10] N. Verma *et al.*, "A micro-power eeg acquisition soc with integrated feature extraction processor for a chronic seizure detection system," *IEEE Journal of Solid-State Circuits*, vol. 45, no. 4, pp. 804–816, 2010.

[11] Y.-H. Chen *et al.*, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 1, pp. 127–138, 2017.

[12] S. Gupta *et al.*, "Deep learning with limited numerical precision," in *Proceedings of The 32nd International Conference on Machine Learning*, 2015, pp. 1737–1746.

[13] B. Moons *et al.*, "Minimum energy quantized neural networks," *arXiv preprint arXiv:1711.00215*, 2017.

[14] K. Hwang and W. Sung, "Fixed-point feedforward deep neural network design using weights+ 1, 0, and- 1," in *Signal Processing Systems (SiPS), 2014 IEEE Workshop on*. IEEE, 2014, pp. 1–6.

[15] B. Moons, B. De Brabandere, L. Van Gool, and M. Verhelst, "Energy-efficient convnets through approximate computing," in *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE, 2016, pp. 1–8.

[16] D. Lin *et al.*, "Fixed point quantization of deep convolutional networks," in *Proceedings of The 33rd International Conference on Machine Learning*, 2016, pp. 2849–2858.

[17] C. Sakr, Y. Kim, and N. Shanbhag, "Analytical guarantees on numerical precision of deep neural networks," in *Proceedings of the 34th International Conference on Machine Learning*, 2017, pp. 3007–3016.

[18] C. Sakr and N. Shanbhag, "An analytical method to determine minimum per-layer precision of deep neural networks," *International Conference on Accoustic, Speech, and Signal Processing (ICASSP)*, 2018.

[19] A. H. Shoeb, "Application of machine learning to epileptic seizure onset detection and treatment," Ph.D. dissertation, Massachusetts Institute of Technology, 2009.

[20] R. G. Andrzejak *et al.*, "Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state," *Physical Review E*, vol. 64, no. 6, p. 061907, 2001.

[21] A. Asuncion and D. Newman, "UCI machine learning repository," 2007.