Efficient AI Predictions through Explainability-driven Neural Network Quantization

Solving increasingly complex real-world problems, continuously contributes to the success of deep neural networks (DNNs) (Schütt et al. 2017; Senior et al. 2020). DNNs have long been established in numerous machine learning tasks and for this have been significantly improved in the past decade. This is often achieved by over-parameterizing models, i.e., their performance is attributed to their growing topology, adding more layers and parameters per layer (Simonyan and Zisserman 2014; He et al. 2016). Processing a very large number of parameters comes at the expense of memory, computational efficiency and invested energy. Such immense storage and energy requirements however contradict the demand for efficient deep learning applications for an increasing number of hardware-constrained devices, e.g., mobile phones, wearable devices, Internet of Things, autonomous vehicles, or robots.

Beyond these, typical applications on such devices, e.g., healthcare monitoring, speech recognition, or autonomous driving, require low latency and/or data privacy. These latter requirements are addressed by executing and running the aforementioned applications directly on the respective devices (also known as “edge computing”) instead of transferring data to third-party cloud providers prior to processing.

Tools for Increasing Network Efficiency

In order to tailor deep learning to resource-constrained hardware, a large research community has emerged in recent years (Deng et al. 2020; Warden and Situnayake 2020). By now, there exists a vast number of tools to reduce the number of operations and model size, as well as tools to reduce the precision of operands and operations (bit width reduction, going from floating point to fixed point). Topics range from neural architecture search (NAS), knowledge distillation, pruning/sparsification, quantization, lossless compression and hardware design.

Quantization and Sparsification

Beyond all, quantization and sparsification are very promising and show great improvements in terms of neural network efficiency optimization (Hoefler et al. 2021; Sze et al. 2017). Sparsification sets less important neurons or weights to zero and quantization reduces parameter's bit widths from default 32-bit float to, e.g., 4 bit integer. These two techniques enable higher computational throughput, memory reduction and skipping of arithmetic operations for zero-valued elements, just to name a few benefits. However, combining both high sparsity and low precision is challenging, especially when relying only on the weight magnitudes as a criterion for the assignment of weights to quantization clusters.

Bit width reduction esopecially has multiple benefits over full precision in terms of memory, latency, power consumption, and chip area efficiency. For instance, a reduction from standard 32-bit precision to 8 bit or 4-bit directly leads to a memory reduction of almost 4x and 8x. Arithmetic with lower bit width is exponentially faster if the hardware supports it. E.g., since the release of NVIDIA's

Turing architecture, 4-bit integer is supported which increases the throughput of the RTX 6000 GPU to 522 TOPS (tera operations per second), when compared to 8-bit integer (261 TOPS) or 32-bit floating point (14.2 TFLOPS (tera floating point operations per second)) (NVIDIA 2018). Furthermore, Horowitz showed that, for a 45nm technology, low-precision logic is significantly more efficient in terms of energy and chip area use (Horowitz 2014). For example, performing 8-bit integer addition and multiplication is 30x and 19x more energy efficient compared to 32-bit floating point addition and multiplication. The respective chip area efficiency is increased by 116x and 27x as compared to 32-bit float. It is also shown that memory reads and writes have the highest energy cost, especially when reading data from external DRAM. This further motivates bit width reduction because it can reduce the number of overall RAM accesses since more data fits into the same caches/registers when having a reduced precision.

Entropy-constrained Quantization

Lowering the entropy of the DNN weights provides benefits in terms of memory as well as computational complexity (Wiedemann et al. 2020; Wiedemann et al. 2021). To recall, the entropy is the theoretical limit of the average number of bits required to represent any element of a distribution (Shannon 1948). Quantization techniques can also reduce the number of bits required to represent weight parameters and/or activations of the full-precision neural network, as they map the respective data values to a finite set of discrete quantization levels (clusters). Providing on such clusters allow to represent each data point in only $\log_2 n$ bit. However, the continuous reduction of the number of clusters generally leads to an increasingly large error and degraded performances. This trade-off is a well-known problem in information theory.

The Entropy-Constrained Quantization (ECQ) algorithm is a clustering algorithm that also takes the entropy of the weight distributions into account.
During quantization, ECQ assigns weight values not only based on their distances to weight centroids found via, e.g., clustering of network parameters, but also based on the information content of the clusters. ECQ is a generalization of the Entropy-Constrained Trained Ternarization (EC2T) (Marban et al. 2020). EC2T trains sparse and ternary DNNs to state-of-the-art accuracies, and its generalization in ECQ allows for the rendition of DNNs with variable bit width.

Explainability-Driven Entropy-Constrained Quantization

Explainable AI (XAI) techniques can be applied to find relevant features in input as well as latent space. Covering large sets of data, identification of relevant and functional model substructures is thus possible. Assuming over-parameterization of DNNs, the authors of (Yeom et al. 2021) exploit this for pruning (of irrelevant filters) to great effect. Their successful implementation shows the potential of applying XAI for the purpose of quantization as well, as sparsification is part of quantization, e.g., by assigning weights to the zero-cluster. Here, XAI opens up the possibility to go beyond regarding model weights as static quantities and to consider the interaction of the model with given (reference) data. The work of (Becking et al. 2021) aims to combine the two orthogonal approaches of ECQ and XAI in order to further improve sparsity and efficiency of DNNs.

For their novel eXplainability-driven Entropy-Constrained Quantization (ECQx), they modify the ECQ assignment function to optimally re-assign the weight clustering based on relevance attributions

from Layer-wise Relevance Propagation (LRP) (Bach et al. 2015) in order to achieve higher performance measures and compression efficiency. The rationale behind using LRP to optimize the ECQ quantization algorithm is two-fold: In terms of an assignment correction, the LRP-generated relevance scores can be used to improve quantization by (1) re-adding "highly relevant" weights, and prevent their quantization to zero, thus upholding model functionality and (2) by assigning functionally irrelevant weights to the zero-cluster, increasing sparsity.

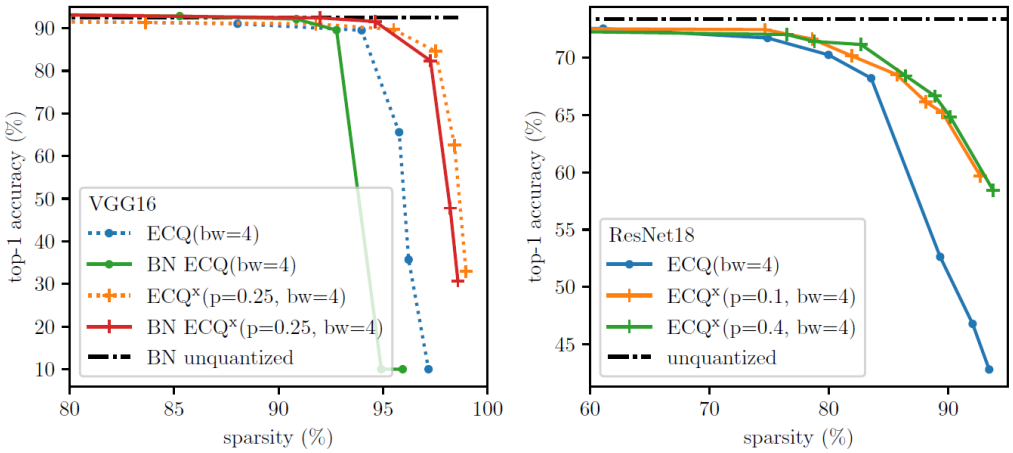

Fig. 1: Sparsity methods comparison.

ECQ^x improves the potential for neural network sparsity beyond the state-of-the-art in low bit network representations, as shown here for VGG networks with and without BatchNorm layers (left), as well as ResNet18 models (right), leading to an up to 103-fold model compression in terms of file size and a considerably reduced energy footprint for using the compressed models for inference. The variable bw indicates the models' bit width, while parameter p controls the per-layer sparsity.

The authors of (Becking et al. 2021) evaluate their explainability-driven network quantization approach on a variety of data domains and neural network architectures, ranging over speech and image categorization data, and from shallow fully connected to deep state-of-the-art DNN architectures.

The comparative results vs. state-of-the-art entropy constrained-only quantization (ECQ) show a performance increase in terms of higher sparsity, as well as a higher compression (see Figure 1). ECQ^x generates low bit width (2-5bit) sparse neural networks while maintaining or even improving model performance. The rendered networks are also highly compressible in terms of file size, e.g., up to 103x compared to the full precision unquantized DNN model, without degrading the model performance. A more detailed discussion of ECQ^x is to be found in the green Open Access article of (Becking et al. 2021) on arxiv.org, which is scheduled to appear as part of *Springer Lecture Notes in Artificial Intelligence: "xxAI – beyond Explainable AI"* in the weeks to come.

Relevance to IToBoS

In IToBoS, we are using DNN models to offer medical experts an assistive AI platform for, e.g., pre-screening and pre-diagnosing large amounts of recorded patients' skin images, or for providing second opinions of AI systems derived from large collections of medical data. With more efficient representations for our powerful AI systems, evaluation bottlenecks can be avoided right from the beginning, while not needlessly expending energy on the AI.

Authors

Sebastian Lapuschkin, Fraunhofer HHI
Frederik Pahde, Fraunhofer HHI

Literature

Schütt, K.T., Arbabzadah, F., Chmiela, S., Müller, K.R., Tkatchenko, A.: Quantumchemical insights from deep tensor neural networks. Nature communications 8(1), 1-8 (2017)

Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Zídek, A., Nelson, A.W.R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D.T., Silver, D., Kavukcuoglu, K., Hassabis, D.: Improved protein structure prediction using potentials from deep learning. Nature 577(7792), 706-710 (2020)

He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770-778 (2016)

Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

Deng, B.L., Li, G., Han, S., Shi, L., Xie, Y.: Model compression and hardware acceleration for neural networks: A comprehensive survey. Proceedings of the IEEE 108(4), 485-532 (2020)

Warden, P., Situnayake, D.: TinyML: Machine Learning with TensorFlow Lite on Arduino and Ultra-low-power Microcontrollers. O'Reilly Media (2020)

Sze, V., Chen, Y., Yang, T., Emer, J.S.: E_cient processing of deep neural networks: A tutorial and survey. Proceedings of the IEEE 105(12), 2295-2329 (2017)

Hoefler, T., Alistarh, D., Ben-Nun, T., Dryden, N., Peste, A.: Sparsity in deep learning: Pruning and growth for e_cient inference and training in neural networks (2021)

NVIDIA Turing GPU Architecture - Graphics Reinvented. Tech. Rep. WP-09183-001 v01, NVIDIA Corporation (2018)

Horowitz, M.: 1.1 computing's energy problem (and what we can do about it). In: 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC).

pp. 10-14 (2014)

Wiedemann, S., Müller, K.R., Samek, W.: Compact and computationally efficient representation of deep neural networks. IEEE Transactions on Neural Networks and Learning Systems 31(3), 772-785 (2020)

Wiedemann, S., Shivapakash, S., Becking, D., Wiedemann, P., Samek, W., Gerfers, F., Wiegand, T.: FantastIC4: A Hardware-Software Co-Design Approach for Efficiently Running 4Bit-Compact Multilayer Perceptrons. IEEE Open Journal of Circuits and Systems 2, 407-419 (2021)

Shannon, C.E.: A mathematical theory of communication. The Bell System Technical Journal 27(3), 379-423 (1948)

Marban, A., Becking, D., Wiedemann, S., Samek, W.: Learning sparse & ternary neural networks with entropy-constrained trained ternarization (ec2t). In: The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 3105-3113 (2020)

Yeom, S.K., Seegerer, P., Lapuschkin, S., Binder, A., Wiedemann, S., Müller, K.R., Samek, W.: Pruning by explaining: A novel criterion for deep neural network pruning. Pattern Recognition p. 107899 (2021)

Becking D., Dreyer M., Samek W., Müller K., Lapuschkin S.: ECQx: Explainability-driven quantization for low-bit and sparse DNNs. arXiv preprint arXiv:2109.04236 (2021)

Back S., Binder A., Montavon G., Klauschen F., Müller K.R., Samek W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS ONE 10(7), e0130140 (2015)