Research Proposal

Data-Free Model Compression and Acceleration Pipeline for Deep Neural Networks

Wanru Zhao, zhaowrenee@gmail.com

Abstract

Quantization and Distillation have emerged as the most promising approaches to compressing and accelerating neural networks. Recent works have proposed data-free compression, aiming to compress models without any real data, which takes great advantage in various senses where real training data cannot be accessed. These methods generate the ’optimal synthetic data’ measuring the batch normalization statistics in each layer to quantize the model. However, compared with the real data, All samples of synthetic data are strictly constrained by BN statistics, causing the significant performance drop of the compressed model. Based on my undergraduate research experiences, this research proposal focuses on my ideas of Learning in ”SCHOOL” and Pre-training Models to deal with the data-free compression and acceleration problem, the topic which I would like to explore during my Ph.D. research.

1 Problem statement

1.1 Background

Deep neural networks (DNNs) have recently achieved great success in many visual recognition tasks. However, existing deep neural network models are computationally expensive and memory intensive, hindering their deployment in devices with low memory resources or applications with strict latency requirements. Therefore, a natural thought is to perform model compression and acceleration in deep networks without significantly decreasing the model performance [4]. The recent techniques for compacting and accelerating DNN models include quantization and knowledge distillation, which are exactly the topics I focused on during my internship at SenseTime Research.

Method	Conference	institution	Compression Type
DFQ [11]	ICCV 2019	Qualcomm	Quant
ZeroQ [2]	CVPR 2020	UC Berkeley	Quant
The Knowledge Within [6]	CVPR 2020	Intel (Habana)	Quant with KD
Deep Inversion [14]	CVPR 2020	NVIDIA	Pruning/KD

Figure 1: Recent techniques for compacting and accelerating DNN models

1. Network quantization compresses the original network by reducing the number of bits required to represent each weight. In the extreme case of the 1-bit representation of each weight, that is binary weight neural networks. The main idea is to directly learn binary weights or activation during model training. There are several works that directly train CNNs with binary weights, for instance, BinaryConnect, BinaryNet and XNOR. For the 8-bit case, which I mainly focused on at SenseTime, where the weight quantization nearly does not affect accuracy, only a small subset of calibration images is needed to determine the activation range in each layer, to re-calibrate batch normalization statistics or to compute layer-wise loss functions to improve quantization performance without fine-tuning.

2. The idea of knowledge distillation (KD) is to compress deep and wide networks into shallower ones, where the compressed model mimicked the function learned by the complex model. The main idea of KD based approaches is to shift knowledge from a large teacher model into a small one by learning the class distributions output via softmax. In SenseTime, Collaborating with various business teams, I promoted model compression via activation-based knowledge distillation in real business scenarios to solve the problem of long-tail distribution in datasets and achieved an improvement of 4% in Accuracy+mA.

1.2 Problem description

With much hardware support lowprecision computations, quantization has emerged as one of the most promising approaches to obtain efficient neural networks. Since the whole training stage is required, quantization aware training methods are considered to be time-consuming and computation intensive. Therefore, posttraining quantization methods were proposed to address this problem, which directly quantizes the FP32 models without retraining or fine-tuning. However, post training quantization methods also require some original training data to calibrate quantized models, which are not often ready-to-use for privacy or security concern, such as medical data and user data.

Recently, it is noticeable that data is an essential requirement in model compression. User data confidentiality protection is becoming a rising challenge in the present deep learning research. In that case, data-free quantization has emerged as a promising method to conduct model compression without the need for user data.

[11] introduces four levels of quantization solutions, in decreasing order of practical applicability in order to distinguish between the differences in applicability of quantization methods. The axes for comparison are whether or not a method requires data, whether or not a method requires error backpropagation on the quantized model, and whether or not a method is generally applicable for any architecture or requires significant model reworking.

- Level 1 No data and no backpropagation required.
- Level 2 Requires data but no backpropagation.
- Level 3 Requires data and backpropagation.
- Level 4 Requires data and backpropagation.

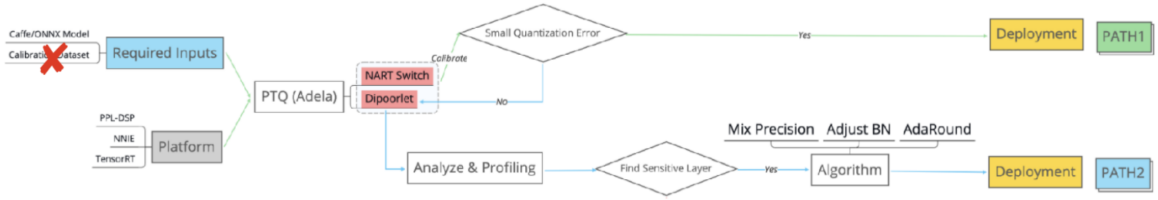

Figure 2: Recent model compression and acceleration pipeline for Deep Neural Networks

My research proposal is based on the definitions above and mainly focuses on Level 1. With no access to data, model quantization naturally becomes less resilient and faces a higher risk of performance degradation. Preliminaries as well as the State-of-the-art (SOTA) and related works are as follows.

1.3 Problem formalization

To obtain quantization parameters of post-training quantization, some data are needed as the input of models to determine the clip values for activations, which we called calibration. And the training/validation images from the application scenario of the model are ideal data to be used to calibrate the models. However, for data-free quantization, we don't have access to any real data. To address this issue, previous works proposed to generate synthetic data based on BN statistics of models. [2] is a typical representation of these works, which proposed to learn an input data distribution that best matches the BN statistics, i.e., the mean and standard deviation, by solving the following optimization problem:

$$\min_{\mathbf{x}^r} \mathcal{L}_{\text{BNS}} = \mathbf{1}^T \mathbf{L}$$

where $\mathcal{L}_{\text{BNS}}$ is the objective function to be minimized, $\mathbf{x}^r$ de- notes the synthetic data composed of the samples x^r, which is initialized by Gaussian random data. 1 is a n -row column vector of all ones, $\mathbf{L} = \{L_0, L_1, \ldots, L_n\}^T$ denotes the BN statistic loss, and the loss of i -th BN layer is represented as:

$$L_i = \|\tilde{\mu}_i^r - \mu_i\|_2^2 + \|\tilde{\sigma}_i^r - \sigma_i\|_2^2$$

where $\tilde{\mu}_i^r$ and $\tilde{\sigma}_i^r$ are the mean and standard deviation statis- tics of the feature distribution of synthetic data at the l -th BN layer, respectively. μ_i and σ_i are the corresponding mean and standard deviation parameters of the l -th BN layer of pre-trained FP32 model.

2 State-of-the-art (SOTA) and related works

1. ZeroQ [2] and the Knowledge Within [6] use distilled datasets to perform data-free quantization, but their methods are model-specific, i.e., one generated dataset can only be used for one model's quantization. Apart from quantization, knowledge distillation is also widely explored.
2. DeepInversion [14] propose a new data-free knowledge distillation framework. It uses the BN statistics variable as an optimization metric to distill the data and obtain high-fidelity

images. BN scheme has also achieved improvements in other tasks. Despite their notable progress, they ignore the true property of real images, i.e. diversity and generalizability on every model.

3 Initial solutions

Inspired by the idea that combining quantization and knowledge distillation, I hope to propose the data-free quantization algorithm to improve both the generalizability and the diversity of the dataset by providing several pre-trained models, which may adopt self-supervised learning benchmarks such as MoCo and SimCLR (the topic I focused on when I was a remote research intern in the University of Notre Dame) to pre-train the model.

3.1 Learning in SCHOOL

Prior works propose to distill fake images by matching the activation distribution given a specific pre-trained model. However, this fake data cannot be applied to other models easily and is optimized by an invariant objective, resulting in the lack of generalizability and diversity whereas these properties can be found in the natural image dataset. To address these problems, I want to design a method capable to generate images suitable for all models by inverting the knowledge in multiple teachers.

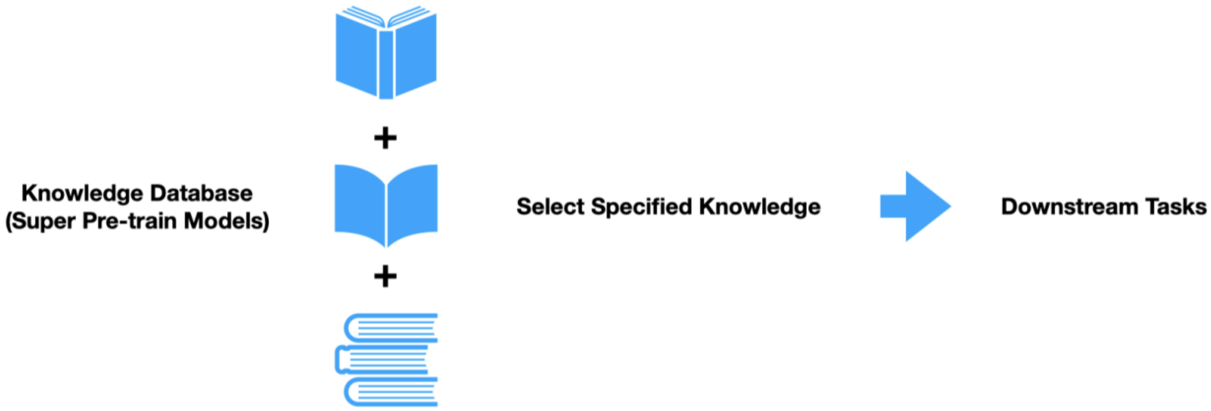

Figure 3: The pipeline incorporates knowledge from teachers and generates a one-for-many synthesized dataset, after which the compressed model can be rapidly deployed to users.

I propose to build a ”school” that contains different aspects of knowledge and distill images with the help of multi-teachers. As a consequence, the data can generalize well to many models. Furthermore, decentralized training and aggregation are to be used to optimize data, therefore the data shares overwhelming diversity and behaves more like natural images with high-fidelity.

3.2 Pre-trained Models

For large-scale datasets in academia and industry, pre-training faces many problems including: There are a lot of data but the current computing power or cluster power cannot support the complete training; the distribution of the data has a large proportion of unlabeled data which can be difficult to use.

The "Pretrained model + Task-related finetuning" mode is widely used. The improvement of the pre-trained model with the same training configuration is likely to lead to the improvement of downstream tasks. The training style of the pre-trained model (supervised or unsupervised) has a greater impact on the downstream tasks.

Self-supervised learning (SSL) defines auxiliary learning tasks that can enhance model's learning capability without requiring any additional annotation effort [10], which can be used to pre-trained models. Recent studies [1,8,9,12,13] are converging on a central concept known as contrastive learning [5]. The results are promising: e.g., Momentum Contrast (MoCo) [7] shows that unsupervised pre-training can surpass its ImageNet-supervised counterpart in multiple detection and segmentation tasks, and SimCLR [3] further reduces the gap in linear classifier performance between unsupervised and supervised pre-training representations.

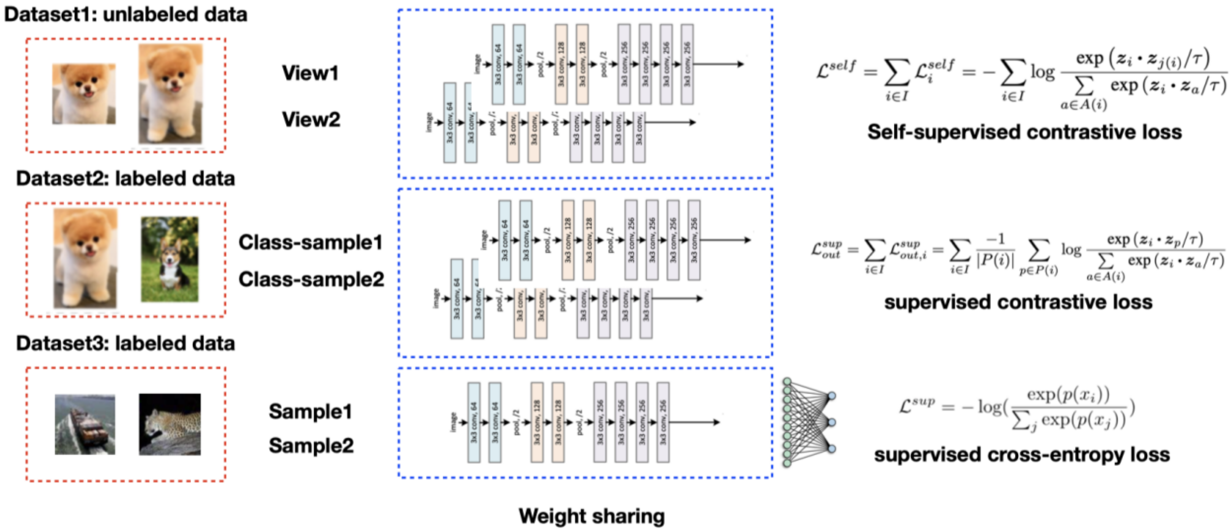
Dataset1: unlabeled data
View1
View2
$\mathcal{L}^{self} = \sum_{i\in I} \mathcal{L}_i^{self} = -\sum_{i\in I} \log \frac{\exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_{j(i)}/\tau)}{\sum_{a\in A(i)} \exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_a/\tau)}$
Self-supervised contrastive loss
Dataset2: labeled data
Class-sample1
Class-sample2
$\mathcal{L}_{out}^{sup} = \sum_{i\in I} \mathcal{L}_{out,i}^{sup} = \sum_{i\in I} \frac{-1}{|P(i)|} \sum_{p\in P(i)} \log \frac{\exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_p/\tau)}{\sum_{a\in A(i)} \exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_a/\tau)}$
supervised contrastive loss
Dataset3: labeled data
Sample1
Sample2
$\mathcal{L}^{sup} = -\log(\frac{\exp(p(x_i))}{\sum_j \exp(p(x_j))})$
supervised cross-entropy loss
Weight sharing

Figure 4: Naive Unified Pre-training Framework

4 Summary and Conclusion

The research will be communicated to the wider community in the way that it can be applied to many common computer vision architectures with a straightforward API call. This is crucial for many practical applications where engineers want to deploy deep learning models trained in FP32 to INT8 hardware without much effort, as well as to protect data privacy of users.

References

[1] Philip Bachman, R. Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *CoRR*, abs/1906.00910, 2019.

[2] Yaohui Cai, Zhewei Yao, Zhen Dong, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Zeroq: A novel zero shot quantization framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13169–13178, 2020.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[4] Yu Cheng, Duo Wang, Pan Zhou, and Tao Zhang. A survey of model compression and acceleration for deep neural networks. *CoRR*, abs/1710.09282, 2017.

[5] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

[6] Matan Haroush, Itay Hubara, Elad Hoffer, and Daniel Soudry. The knowledge within: Methods for data-free model compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8494–8502, 2020.

[7] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *CoRR*, abs/1911.05722, 2019.

[8] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, pages 4182–4192. PMLR, 2020.

[9] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

[10] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Zhaoyu Wang, Li Mian, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *arXiv preprint arXiv:2006.08218*, 1(2), 2020.

[11] Markus Nagel, Mart van Baalen, Tijmen Blankevoort, and Max Welling. Data-free quantization through weight equalization and bias correction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1325–1334, 2019.

[12] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *CoRR*, abs/1906.05849, 2019.

[13] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018.

[14] Hongxu Yin, Pavlo Molchanov, Zhizhong Li, Jose M. Alvarez, Arun Mallya, Derek Hoiem, Niraj K. Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via deepinversion. *CoRR*, abs/1912.08795, 2019.