

Data-Free Network Quantization With Adversarial Knowledge Distillation

Yoojin Choi¹, Jihwan Choi², Mostafa El-Khamy¹, Jungwon Lee¹

¹SoC R&D, Samsung Semiconductor Inc., San Diego, CA ²DGIST, Korea

{yoojin.c, mostafa.e, jungwon2.lee}@samsung.com jihchoi@dgist.ac.kr

Abstract

Network quantization is an essential procedure in deep learning for development of efficient fixed-point inference models on mobile or edge platforms. However, as datasets grow larger and privacy regulations become stricter, data sharing for model compression gets more difficult and restricted. In this paper, we consider data-free network quantization with synthetic data. The synthetic data are generated from a generator, while no data are used in training the generator and in quantization. To this end, we propose data-free adversarial knowledge distillation, which minimizes the maximum distance between the outputs of the teacher and the (quantized) student for any adversarial samples from a generator. To generate adversarial samples similar to the original data, we additionally propose matching statistics from the batch normalization layers for generated data and the original data in the teacher. Furthermore, we show the gain of producing diverse adversarial samples by using multiple generators and multiple students. Our experiments show the state-of-the-art data-free model compression and quantization results for (wide) residual networks and MobileNet on SVHN, CIFAR-10, CIFAR-100, and Tiny-ImageNet datasets. The accuracy losses compared to using the original datasets are shown to be very minimal.

1. Introduction

Deep learning is now leading many performance breakthroughs in various computer vision tasks [1]. The state-of-the-art performance of deep learning came with over-parameterized deep neural networks, which enable extracting useful representations (features) of the data automatically for a target task, when trained on a very large dataset. The optimization framework of deep neural networks with stochastic gradient descent has become very fast and efficient recently with the backpropagation technique [2, Section 6.5], using hardware units specialized for matrix/tensor computations such as graphical processing units (GPUs).

Figure 1: Data-free adversarial knowledge distillation. We minimize the maximum of the Kullback-Leibler (KL) divergence between the teacher and student outputs. In the maximization step for training the generator to produce adversarial images, the generator is constrained to produce synthetic images similar to the original data by matching the statistics from the batch normalization layers of the teacher.

The benefit of over-parameterization is empirically shown to be the key factor of the great success of deep learning, but once we find a well-trained high-accuracy model, its deployment on various inference platforms faces different requirements and challenges [3, 4]. In particular, to deploy pre-trained models on resource-limited platforms such as mobile or edge devices, computational costs and memory requirements are the critical factors that need to be considered carefully for efficient inference. Hence, model compression, also called network compression, is an important procedure for development of efficient inference models.

Model compression includes various methods such as (1) weight pruning, (2) network quantization, and (3) distillation to a network with a more efficient architecture. Weight pruning and network quantization reduce the computational cost as well as the storage/memory size, without altering the network architecture. Weight pruning compresses a model by removing redundant weights completely from it, i.e., by setting them to be zero, so we can skip computation as well as memorization for the pruned weights [5–12]. Net-

Work done when the author was with Samsung as a visiting scholar.

