Offline Response-based Knowledge Distillation in Convolutional Neural Networks

Srinivas Rao Daru
ECE, UCSD
sdaru@ucsd.edu
A59003596

Sri Harsha Pamidi
ECE, UCSD
spamidi@ucsd.edu
A59005361

Venkata Sai Abhiram Medisetti
ECE, UCSD
vmediset@ucsd.edu
A59005405

Neeharika Gonuguntla
CSE, UCSD
ngonugun@ucsd.edu
A59004781

Pujika Kumar
ECE, UCSD
pukumar@ucsd.edu
A59005004

Abstract

Deep neural networks are proven to be successful in both industry and research in recent years, particularly for computer vision problems. Deep learning's enormous success can be attributed to its capacity to encode vast amounts of data and manipulate billions of model parameters. However, not only because of the high computational complexity, but also because of the massive storage needs, it is difficult to install these cumbersome deep models on devices with low capabilities, such as mobile phones and embedded devices. A number of model compression and acceleration approaches have been developed to this purpose. Knowledge distillation is a sort of model compression and acceleration that efficiently learns a small student model from a large teacher model. In this project we aim to implement offline response-based knowledge distillation from a pre-trained ResNet50 or Inception_v3 network to a small student CNN network. We aim to check if the inception net can be shortened to a smaller model and achieve comparable performance.

1 Introduction

Knowledge Distillation is used as a model compression technique in which a small model is taught to imitate a larger model (or ensemble of models) that has already been trained. The larger model is the teacher, and the smaller model is the student. This training scenario is commonly referred to as "teacher-student" model. While large models (such as very deep neural networks or ensembles of numerous models) have greater knowledge capacity than small models, this capacity may be underutilized. Even if a model only uses a small portion of its knowledge capacity, evaluating it can be computationally costly. The process of shifting knowledge from a large model to a smaller one while keeping validity is known as knowledge distillation. Because smaller models are less expensive to infer, they can be run on less powerful hardware (such as a mobile device). Knowledge distillation has been effectively used in a variety of machine learning applications, including object recognition.In [1], model compression is presented as a way of transferring knowledge from a large model or an ensemble of models into training a small model without sacrificing accuracy. In [3], Knowledge distillation is referred to as the process of learning a tiny model from a larger one. A large instructor model supervises a tiny student model in knowledge distillation. The basic premise is that the student model is similar to that of the teacher. Response-based knowledge is typically defined as the neural response of the teacher model's final output layer. The main idea is to directly mimic the

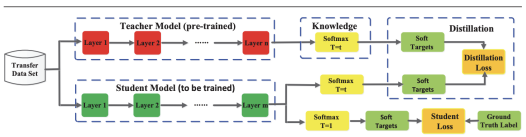

Figure 1: Response based knowledge distillation[3]

final prediction of the teacher model. Response-based knowledge distillation is a simple yet effective method for model compression that has been widely used in a variety of tasks and applications[2].

2 Dataset

The CIFAR-10 dataset, introduced in [4] contains 60000 32x32 color images divided into 10 classes, each with 6000 images. There are 50000 training images and ten thousand test images. The dataset is divided into five training batches and one test batch, with each batch containing 10,000 photos. The test batch contains exactly 1000 photographs from each class that were chosen at random. The remaining images are distributed in random order in the training batches, however certain training batches may contain more images from one class than another. The training batches each include exactly 5000 photos from each class.

3 Idea

We plan to compress the ResNet50 and Inception-v3 model which is a 48-layer deep pre-trained convolutional neural network model using response-based knowledge distillation. The Inception-v3 is a trained on over a million photos from the ImageNet collection. We try to build a correspondence between the student and teacher networks by comparing the responses of the final layer soft outputs of the student and teacher networks using the Kullback-Leibler divergence loss described in this paper [2] and as shown in Figure-1.

There exists many ways in which we do such knowledge distillation depending on the way knowledge is transferred from the teacher to student. One of such ways is Offline Distillation, where the teacher model is already pretrained. This is beneficial as the offline distillation involves knowledge transfer in one way, which is only from teacher to student (teacher doesn't get updated). Since the teacher is already pretrained, we expect the model to learn faster.

4 Model

We plan to explore various student network combinations of convolutional layer and fully connected layer architectures of the student network, which are able imitate the teacher network and get a satisfactory performance on the test data. we plan to use both the KL divergence loss between the outputs from the student and teacher network and also the cross entropy loss between the ground truth label and the soft logits of the student model. This enables the student network to learn the task at hand and simultaneously learn to replicate the behavior of the teacher network.

References

[1] C. Bucilua, R. Caruana, and A. Niculescu-Mizil. Model compression, in proceedings of the 12 th acm sigkdd international conference on knowledge discovery and data mining. *New York, NY, USA*, 2006.

[2] J. Gou, B. Yu, S. J. Maybank, and D. Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, Mar 2021.

[3] G. Hinton, O. Vinyals, J. Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

[4] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.