Knowledge Distillation Stanford40 and CIFAR100

Mohammad Zafari

April 15, 2025

Contents

1 Introduction

This project use the concepts of Knowledge Distillation via Attention-based Feature Matching. Knowledge distillation is the technique for transferring knowledge from a source neural network to a target neural network. The source network, referred to as a teacher, indicates a large network that is highly regularized via pre-training, and the target network, referred to as a student, is a smaller network for a specific task.

2 Requirements

First, the Stanford40 data must be formatted in a way that is suitable for use in:
torch.utils.data.DataLoader

We use the following file to do this:
Stanford40DataLoader

The CIFAR100 dataset can also be used as follows:
torchvision.datasets.CIFAR100

3 Networks

It this training 2 diffrent networks has been used:

In main4.py for stanford40 data we have:
Teacher
resnet34im = "ResNet-34 model from Deep Residual Learning for Image Recognition

Student
resnet18im = "ResNet-18 model from Deep Residual Learning for Image Recognition

In others:
Teacher
WRN40X2 = WideResNet(depth=40, Widen _factor=2)

Student
WRN40X2 = WideResNet(depth=16, Widen_factor=2)

4 Parameters

The parameter sizes before and after using knowledge distillation are listed in the table below.

DATA	CIFAR100	STANFORD40
TEACHER	2,255,156	2,255,156
STUDENT	703,284	695,544

5 Results

TEACHER	WRN40X2	WRN40X2
STUDENT	WRN16X2	WRN16X2
DATA	CIFAR100	Stanford40
TEACHER	0.7620	—
STUDENT	0.7289	—
STUDENT(5 times mean)	0.7547	0.3693
EXPERIMENT ACCURACY	0.7556	0.2713
EACH EPOCH TIME(S)	60	164
TOTAL EPOCH	240	65

6 References

1. Show, Attend and Distill:Knowledge Distillation via Attention-based Feature Matching
2. Attention-feature-distillation
3. Stanford40DataLoader