EEE7331-01

SPECIAL TOPICS IN DEEP LEARNING

2022 AUTUMN

Assignment 2

Due Date: Before 13 December (Tuesday) Midnight to LEARNINGUS.

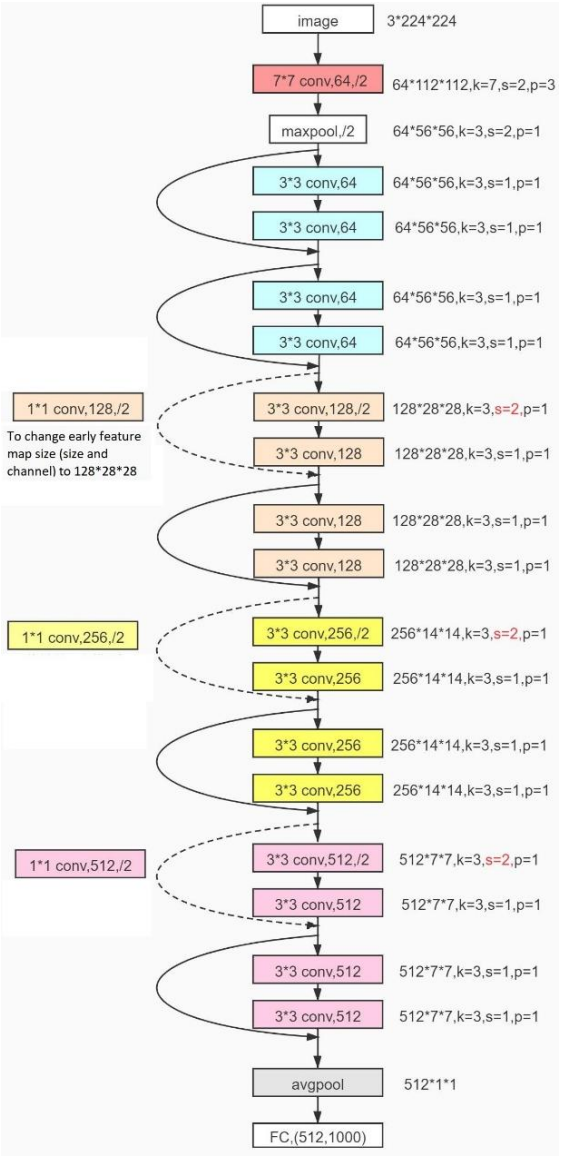
INSTRUCTIONS:

1. This question paper consists of 6 pages.
2. Attempt ALL questions.
3. This is an individual-based assignment. Discussions among students are welcome. However, any cheating attempt will be penalized.
4. Attach your code as an appendix. You may show the **code snippet** (not the whole code) to **aid your explanation**.
5. Upload the solution in a single pdf file.
6. Use English in the report.

Question 1: Knowledge Distillation (KD) (50%)

Part A: Setup a Teacher Network

We will use CIFAR-10 Dataset (https://www.cs.toronto.edu/~kriz/cifar.html) again. The teacher network is an ImageNet prebuilt ResNet18 [1]-[3] fined tuned with CIFAR 10. The following figure is the architecture of ResNet18.



Layer name	output size	18-layers	No. of Blocks
conv1	112x112	7x7,64, stride 2	
conv2_x	56x56	3x3 max pool, stride 2	
		3x3,64	x 2
		3x3,64	
conv3_x	28x28	3x3,128	x 2
		3x3,128	
conv4_x	14x14	3x3,256	x 2
		3x3,256	
conv5_x	7x7	3x3,512	x 2
		3x3,512	
	1x1	average pool, 100-d fc, softmax	

Figure 1: ResNet18 Architecture

Note the following:

- Change the softmax layer to 10 neurons, corresponding to 10 classes of CIFAR-10.

- Use **original CIFAR 10 images (32 x 32 x 3) to input ResNet18**. Modify the stride and/or padding to make the last feature map size **8x8x512**. You should not modify the network architecture, i.e., 17 convolution layers and 1 max pooling layer and filter size.
  Detail what you have changed to make this happen.
- **Retrain** softmax layer and **finetune the rest in ResNet18** with CIFAR 10.

(1) Describe training setup, e.g., data pre-processing/augmentation, initialization, hyperparameters such as learning rate schedule, momentum, dropout rate, batch number, etc. Present them systematically in **table form**.

(2) Plot the following:
- Training and test loss (not classification accuracy) vs. epoch.
- Classification accuracy on the training and test set vs. epoch.

(4) Fill the table with the final accuracy that you obtained.

Training Accuracy (%)	**Testing Accuracy (%)**

Remark: The testing accuracy should be at least 90% and above. With proper tuning of the network hyperparameters, you can achieve more than 95%. A large gap in training and testing accuracies is not allowed. The score for this part is proportional to the accuracy you can obtain. Note that the accuracy values should be consistent with the graph shown in (3). The answer in (4) is discounted if no graph is shown.

Part B: Setup a Student Network and perform Knowledge Distillation

Create a student network with the following specifications:

	Stage 1		Stage 2		Stage 3		Stage 4	Stage 5
Layer	Conv	Pool	Conv	Pool	Conv	Pool	FC	Output
	$(5,32)_{/1,1}$ RELU	$2_{/1,0}$	$(3,64)_{/1,1}$ RELU	$2_{/1,0}$	$(3,128)_{/1,1}$ RELU	$2_{/1,0}$	500	10, softmax

where $(m, n)_{/a,b}$ with n filters of size $m \times m$, where the stride and padding are a and b, respectively. $p_{/r,s}$ denotes the max-pooling layers with a window of $p \times p$, where the stride and padding are r and s, respectively.

(1) Generate a soft target using the teacher network developed in part A. The temperature T depends on the scenarios specified in the table below.

(2) Describe the training setup for the student network. Then, apply the same tricks (data augmentation, BN, dropout, etc., if any) as in the teacher network in part A.

(3) Train the student network with the loss function specified in Lecture 7, i.e., a mix of CEs made by the soft and hard target. Set **λ=0.1 by default** (unless stated otherwise). Plot the following:

(a) Training and test loss vs. epoch.

(b) Classification accuracy on the test set vs. epoch.

(4) Fill in the table below:

Model	**Temperature, *T***	**Training Accuracy (%)**	**Testing Accuracy (%)**
Teacher Network (Part A)	-		
Student Network ($\lambda = 0$)	-		
Student Network ($\lambda = 1$)	20		
KD 1	5		
KD 2	10		
KD 3	20		

Note that during *testing*, *T* is set to 1 in the softmax function regardless of *T* used in the Teacher Network and training.

Discuss your observation from the experiment results.

(5) (Optional for a bonus point) You may find configurations above may not be sufficient to let the student network performs comparably to the teacher network. Try more *T* and adjust λ to achieve the goal of KD.

Question 2: Continuous Learning (50%)

(Before solving this problem, it is advised to go through [4])

Now we want to add a new task – face recognition on top of the ResNet18 developed in Qs 1 using Feature extraction, Fine Tuning, and LWF discussed in Lecture 7. The goal is to evaluate whether they can learn a new task (face recognition) while preserving performance on old tasks (CIFAR10 classification).

The face dataset consists of 100 subjects, with 50 samples per subject for the new task.

(1) **Feature Extraction: Re-train the last (softmax) layer** of that developed in question 1 with face training samples (**40 samples per subject**). This is equivalent to you freezing the rest of the layers in the network and replacing them with a new classifier that tunes to the face dataset. Note number of output nodes should change to 100 instead of 10. Evaluate the model with a face test set composed of **10 samples per subject**.

(2) **Fine Tuning:** Fine-tune a few chosen convolution and FC layers and re-train the softmax classifier on top with face training images. Fill in the table below. Then, evaluate the model with the face test set.

Conv layers that you choose to fine-tune	**Learning Rate (specify in this form: x% of the original learning rate)**	**Remark (if any)**
Conv 5-3 (Example)		

(3) **Learning Without Forgetting (LFW):** Describe what you do to implement the LFW in point form. Detail the parameters used in the experiments such as learning rate (a% of original learning rate), temperature T and λ_0. The L2 regularizer coefficient is set to 0.0005. You may refer to [4] for reference.

For every method,

(a) Show both training and test loss vs. epoch figures.

(b) Show both classification accuracy on the training and test set vs. epoch figures.

(4) Fill in the table

		Training Accuracy (%)	**Testing Accuracy (%)**
Feature Extraction	**Old Task***		
	New Task		
Fine Tuning	**Old Task***		
	New Task		
LFW	**Old Task***		
	New Task		

* Old task performance refers to the performance of CIFAR 10 image classification **after** the network has been modified for face recognition. The goal is to see how much the network preserves its performance on the old task after tuning it for the new task.

Discuss the experiment's result and your observations.

[1] https://pytorch.org/hub/pytorch_vision_resnet/ (Pytorch)
[2] https://kr.mathworks.com/help/deeplearning/ref/resnet18.html (MATLAB)
[3] https://github.com/qubvel/classification_models (Keras and TensorFlow Keras)
[4] Z. Li and D. Hoiem, "Learning without Forgetting," IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–1, 2017, doi: 10.1109/TPAMI.2017.2773081.