Homework 6
601.482/682 Deep Learning
Fall 2023

Oct 18, 2023

Due 11:59pm on Nov 8, 2023

Please submit 1) a single zip file containing your Jupyter Notebook and PDF of your Jupyter Notebook to "Homework 6 - Notebook" and 2) your written report (LaTeX generated PDF) to "Homework 6 - Report" on Gradescope (Entry Code: BBVDNN)

Important: You must program this homework using the PyTorch framework. We highly recommend using Google Colaboratory.

Important: If you don't have local GPU access, you should port the provided Python scripts to Colaboratory and enable GPU in that environment (under Edit->Notebook Settings). Training should converge in less than 30 min. If your model does not make significant updates in that time, you should re-examine your code. Either way, this is a reminder to start the assignment early.

1. *Unsupervised Pre-training.* In this problem, you will attempt the 2017 Endoscopic Instrument Challenge.[1] You are given a pre-processed dataset consisting of endoscopic frame images (*not* in sequential order). The goal is to train a network which takes each RGB frame as an input and predicts a pixel-wise segmentation mask that labels the target instrument type and background tissue. Additionally, we introduce an unsupervised pre-training method and compare the performance of training on a small labeled dataset with/without pre-training. This is relevant for real-life medical image problems, where there is usually a shortage of data labels.

   Data Folder We have provided a well-structured dataset. It consists of '/segmentation' and '/colorization'. In each sub-folder, there are '/train' and '/validation' for training purposes.

   The **main goals** of the homework are as follows. Concrete TODOs are enumerated on the next page.

   - *Complete the Network structure for segmentation task (a-c).* The network structure we provide is a simplified U-Net, which is a very popular framework in medical image segmentation task. Read and understand the code in `unet.py`, the implementation is missing the last layer and the last activation function. Next, fill in the missing components for 1(a). For the segmentation task, train ONLY with the frames in '/segmentation/train' and validate and test with the frames in '/segmentation/validation' and '/segmentation/test' respectively. The original input image is a $256 \times 320 \times 3$ RGB image. The ground truth label is a grey-scale image that has the same dimension, where different gray values indicate the instrument type or background tissue.
   - *Pre-training by self-supervised colorization (d).* Image colorization training[2] is a common way of self-supervised learning.[3] The idea is to take grey-scale images as an input and predict colorized images, similar to filling in colors in a draft painting. You must

[1] `https://endovissub2017-roboticinstrumentsegmentation.grand-challenge.org`

[2] Larsson, G., Maire, M., & Shakhnarovich, G. (2017). Colorization as a proxy task for visual understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6874-6883).

[3] Jing, L., & Tian, Y. (2020). Self-supervised visual feature learning with deep neural networks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence.

use the **same** U-Net structure as above to train a model for this colorization task. Then use the pretrained weights as initialization for the segmentation task.

For the colorization task, you need to train on '/colorization/train_cor' and validate on '/colorization/validation_cor'. You are given `mapping.json` that contains the label of each grey level. We have also provided grey-scale image for each input in each subfolder of '/colorization' for your colorization task input.

Now that you know the broad goals and data sets, please complete the following TODOs.

(a) Train a segmentation network using the frames in the '/segmentation/train' folder. Please complete the DICE score function to evaluate your model, and write from scratch a DICE loss function as your network loss[4]. (*Hint: You need to convert the grey-scale label mask to one-hot encoding of the label and then calculate the DICE score for each label*). Please train the network until convergence (should take around 30 min) using the default provided hyperparameters and provide a figure of training loss and validation loss w.r.t. epochs (in a single figure). Please report your performance (DICE score) on the test dataset, you should expect a DICE score > 0.5. (*Hint: Using BatchNorm might help you achieve better performance.*)

(b) Introduce meaningful data augmentation (e.g. vertical and horizontal flips) and train the network until convergence using the same hyperparameters as (a). Please plot the training loss and validation loss on a single figure again and report test dataset performance, you should expect a DICE score > 0.6.

(c) Train on the colorization task using frames from the '/colorization/train_cor' folder. Use hyperparameters that seem reasonable (based on your previous experiments) and mean squared error as your loss function. Please provide a figure of training loss w.r.t. epochs until your model converges. Then save your model to initialize the network for the next task.

(d) Load the colorization pre-trained model and start training for the segmentation task using the frames in the '/segmentation/train' folder. Make sure you are using the same hyperparameters as you did in the former task, and please clearly state them in your report. Plot the figure of training loss and validation loss. Report test dataset performance. Do you see a difference with the former result in (b)? (*Hint: Since this is a relatively simple dataset, you might not actually observe differences in performance.*)

2. *Transfer Learning*. Please download the fashion MNIST dataset [5] as used in HW4 and download the VGG16 model (`https://pytorch.org/docs/stable/torchvision/models.html`).

(a) Randomly initialize all parameters in VGG16 and try to train your model to learn the Fashion MNIST classification task. What's the accuracy you achieve? Please report your test accuracy on the test dataset. You should expect an accuracy $> 85\%$.

(b) Load the pre-trained VGG16 model from torch vision models. Freeze all but the last layer: randomly initialize the last layer of your network and fine-tune this. What accuracy do you get now? Please again report your test accuracy on the test dataset. You should expect an accuracy $> 60\%$.

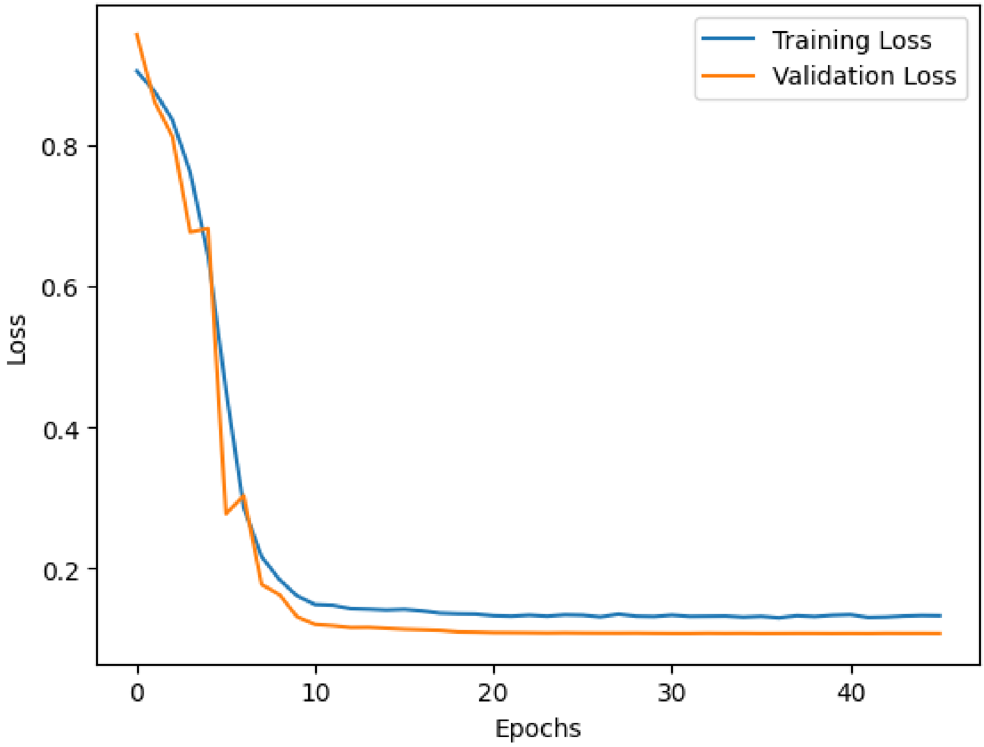
(c) Now, imagine a scenario in which you want to train the VGG16 model on an entirely new dataset and will fine-tune either the model from (2a) or (2b). Which pre-trained model is the preferred starting point for your new use case?

[4] Read more about the DICE score: `https://medium.com/datadriveninvestor/deep-learning-in-medical-imaging-3c1008431aaf`

[5] The full FashionMNIST dataset can be downloaded from the official website here: `https://github.com/zalandoresearch/fashion-mnist`

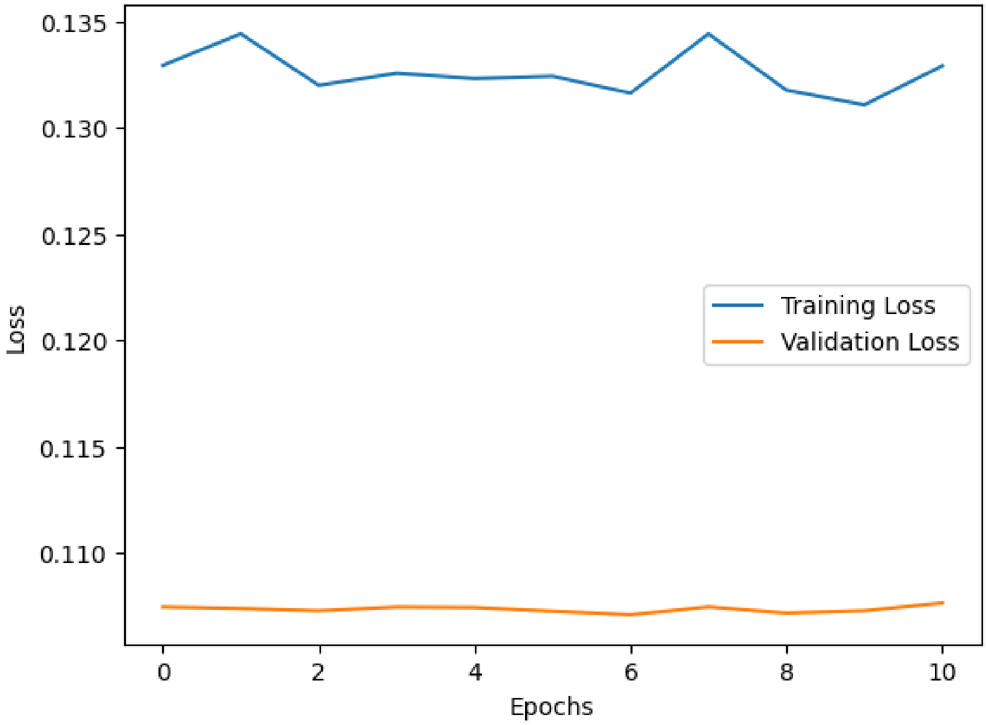
Answers

1. (a) Test DICE Score: 0.5295



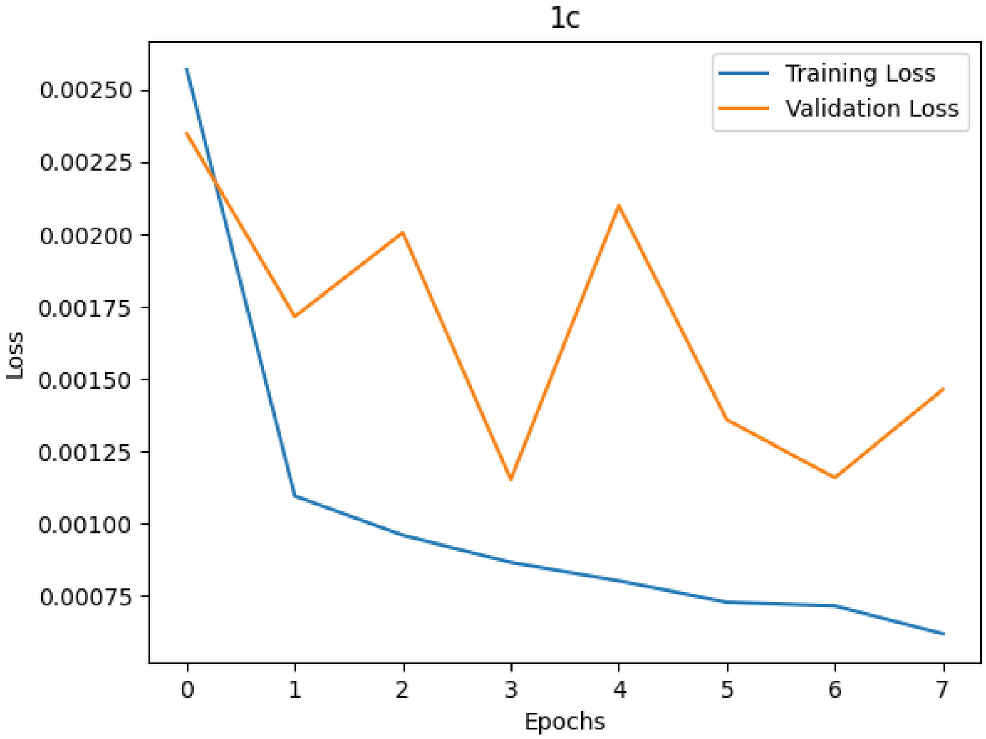
Early stopping occurred at epoch 46 due to no improvement in validation loss. I used a patience of 4 epochs for all early stopping in this assignment.

(b) Test DICE Score: 0.5295



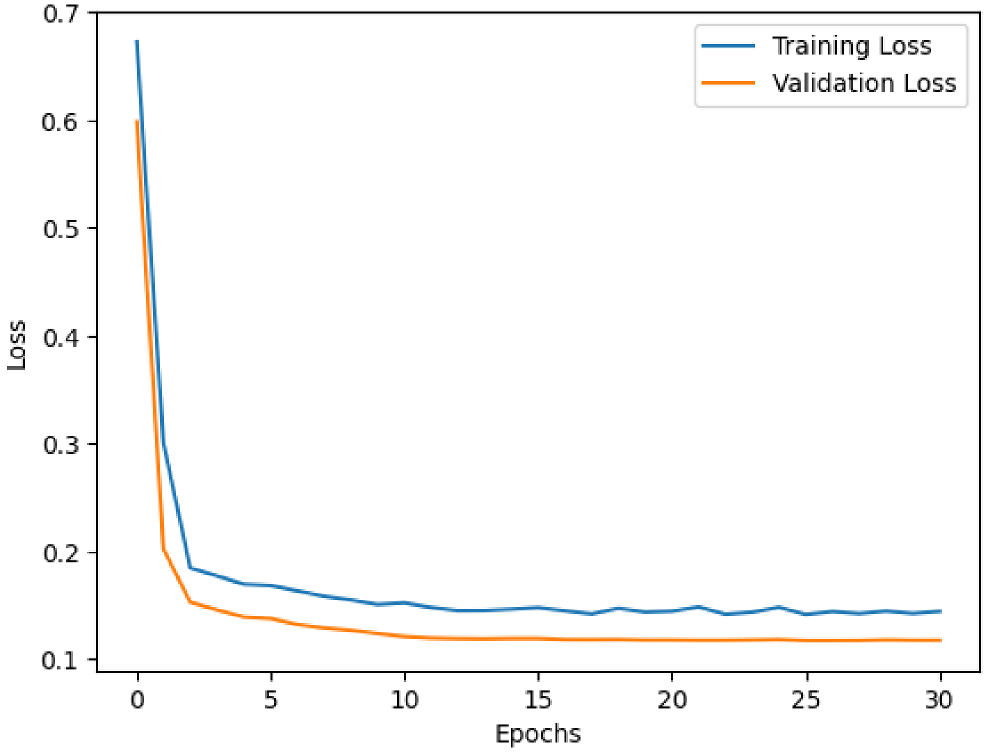
Early stopping occurred at epoch 11. Data augmentation (horizontal flip, vertical flip, and 30 degree rotation) appears to have had no effect on the DICE score, perhaps because this is a relatively simple dataset.

(c) Unlike in parts a and b, where the model had 6 classes to account for the 6 possible categories for segmentation, this new colorized model was only given 3 classes. This is so the output images have 3 channels in accordance with what is needed for part d. 200 was kept as the number of epochs, with early stopping implemented just as it was for parts a and b. I also kept the same learning rate (0.001), Adam optimizer, and learning rate scheduler from parts a and b.



Early stopping occurred at epoch 8.

(d) Test DICE Score: 0.6991



Early stopping occurred at epoch 31. I used the same hyperparameters as part c, which are the same ones as used in parts a and b. I also included the same data augmentation as part b (horizontal flip, vertical flip, and 30 degree rotation). There is a marked increase in the test DICE score between 1b and 1d, suggesting that colorization prior to segmentation can increase the accuracy of the network since nothing else differs between the two.

2. (a) I trained my model for 10 epochs (took almost 3 hours on T4 GPU) and got a test accuracy of 90.89%.

   (b) I trained this model for 10 epochs as well and got a test accuracy of 85.62%.

   (c) In general, a network that has been trained on a larger set or multiple sets of data is preferred, meaning 2b is the better option. It can utilize the knowledge gained from pre-training on ImageNet and fine-tuning to the Fashion MNIST dataset, whereas 2a would lack that broader knowledge from ImageNet since it was not pre-trained on it. Therefore, 2b should yield faster convergence and better generalization through transfer learning.