

E0 270 : Machine Learning

Assignment 2

Due: May 26, 2021

Instructions:

- There are 2 questions
- Create separate submission folder for both problems
- Compress your folder containing files to a zip file. Then rename .zip extension to .pdf and upload.

1. The task is to implement convolutional neural network for image classification. In this problem you will be using Fashion-MNIST dataset. The task is to recognize the image and identify it as one of 10 classes.

Dataset Fashion-MNIST is a dataset of Zalando’s article images—consisting of a training set of 60,000 examples and a test set of 10,000 examples. Each example is a 28x28 grayscale image, associated with a label from 10 classes.

Each image is 28 pixels in height and 28 pixels in width, for a total of 784 pixels in total. Each pixel has a single pixel-value associated with it, indicating the lightness or darkness of that pixel, with higher numbers meaning darker. This pixel-value is an integer between 0 and 255. The training and test data sets have 785 columns. The first column consists of the class labels (see above), and represents the article of clothing. The rest of the columns contain the pixel-values of the associated image.

Label	Description
0	T-shirt/top
1	Trouser
2	Pullover
3	Dress
4	Coat
5	Sandal
6	Shirt
7	Sneaker
8	Bag
9	Ankle boot

Table 1: Labels for Fashion-MNIST dataset

You can load Fashion MNIST dataset using pytorch/tensorflow API. For more information of this dataset see, <https://github.com/zalandoresearch/fashion-mnist>

Plan of Action This assignment has to be coded using pytorch or tensorflow. In this assignment, you need to do the following:

- Train a convolutional neural network using Fashion-MNIST dataset.
- Validate classification performance of your model using validation set. Tune your hyper parameters using validation set.
- After hyper parameter tuning is completed, fix your hyper-parameters and model parameter and test your model’s performance on the testing set

Deliverables You need to create a folder which contains the following. The names of folder and files should be self-explanatory:

- Prepare a report. Please remember that report is an important component for evaluation.
- The python code for training the model. The training code should generate a plot of training loss vs number of epochs while training
- You need to save the trained model in a specific format. During test time, your code should use this saved model and not retrain it from scratch.
- The python code for testing the model. On running this file, your model should generate the classes for all the images in the test data (included with the dataset API) and output them to files titled cnn.txt. The test code should also give accuracy measured as $Accuracy = \frac{N_c}{N}$, where N_c is the number of corrected classified data samples, and N is the total number of samples and, also confusion matrix.

In your report, you need to briefly describe what you have done, present the results and provide a brief discussion of the results you obtained. Please feel free to play with the model architecture and hyper-parameter settings. Please make it concise and to-the-point. Your report should also include the following evaluation metrics

- (a) Graph of training loss vs number of epochs while training
 - (b) Classification accuracy: $Accuracy = \frac{N_c}{N}$, where N_c is the number of corrected classified data samples, and N is the total number of samples of the validation set.
 - (c) Confusion matrix
2. Use an autoencoder for learning unsupervised representations of a text. For autoencoder architecture, use a bi-directional LSTM for the encoder and unidirectional LSTM for the decoder. For both of them, keep 100 as word embeddings and hidden state dimension, and the rest of the parameters in default settings. We will use the Yelp restaurant review dataset for this task, and it is in the assignment folder. Use the Yelp dataset files train.txt and valid.txt for training and validation of the autoencoder. Then do the following tasks and report your observations. (Refer to this tutorial to learn how to build a seq-2seq model https://pytorch.org/tutorials/intermediate/seq2seq_translation_tutorial.html)
- (a) Plot the validation loss per epoch.
 - (b) Find the reconstruction loss (remove the loss due to padding while reporting score) and BLEU score on test.txt file. (Use this package <https://pypi.org/project/bleu/> for BLEU score)
 - (c) Show the effect teacher forcing on the above metrics using a graph.
 - (d) Use the learned model to generate unsupervised representation on the sentiment analysis dataset in the yelp folder. Do a dimensionality reduction and visualize it using a 2-D scatter plot along with their labels. (Use files 1000.pos and 1000.neg)
 - (e) Train a logistic regression classifier on the representations from the earlier task and report the classification accuracy on files 100.pos and 100.neg.
 - (f) Compare the earlier performance with LSTM based text classifier trained 1000.pos and 1000.neg files with the same architecture as the encoder used in the autoencoder.

The report should include details of training, methodology of experiments, and observations along with figures.