DATA 255 Deep Learning Technologies
Homework -3
Deadline: November 13, 2023 (Midnight 11.59 PM)
20 points

Problem 1 (8+2 = 10 pts): Apply Variational Autoencoder on the **Fashion MNIST** Dataset.

a. Use minimum of 3 convolutional layers in the encoder and 3 deconvolutional layers (Conv2DTranspose/ upscale) in the decoder.
b. Display how the latent space clusters different classes of the training data.

Problem 2 (2+ 8 = 10 pts): Use the IMDB Movie review dataset:

a. Perform Text Preprocessing
    a. Tokenization
    b. Stopwords removing
    c. HTML removing
    d. Convert to lower case
    e. Lemmatization/stemming
b. Build the following sentiment analysis models and create a performance comparison table:
    a. TF-IDF + GausianNB
    b. Word2Vec (CBoW) + GausianNB
    c. Glove + GausianNB

BONUS (2 pts): Write in your own words what is Byte Pair Encoding (BPE) and mentioned the steps involved in BPE tokenization. Apply Byte Pair Encoding (BPE) for 5%, 10%, 15% and 20% of the IMDB training dataset. Compare the BPE in terms of number of generated tokens for the varying datasets.

Useful link-

Data Download:

Fmnist: keras- https://keras.io/api/datasets/fashion_mnist/

Keras example - https://keras.io/examples/generative/vae/

BPE resources -

Huggingface BPE - https://huggingface.co/learn/nlp-course/chapter6/5?fw=pt

https://www.geeksforgeeks.org/byte-pair-encoding-bpe-in-nlp/

https://towardsdatascience.com/training-bpe-wordpiece-and-unigram-tokenizers-from-scratch-using-hugging-face-3dd174850713

CIFAR10: Keras: https://keras.io/api/datasets/cifar10/

Pytorch: https://pytorch.org/vision/stable/generated/torchvision.datasets.CIFAR10.html

Movie review: https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews

Gensim- https://radimrehurek.com/gensim/models/word2vec.html

You are required to submit:

1. An MS/PDF/Scanned document:
    a. Include all the steps of your calculations.
    b. Attach screenshots of the code output.
    c. Include the summary of the model
    d. Include a Table - Mention all the hyperparameters you selected: activation function in hidden layer and output layer, weight initializer, number of hidden layers, neurons in hidden layers, loss function, optimizer, number of epochs, batch size, learning rate, evaluation metric
2. Source code:
    a. Python (Jupyter Notebook)
    b. Ensure it is well-organized with comments and proper indentation.

- Failure to submit the source code will result in a deduction of 5 points.
- Format your filenames as follows: "your_last_name_HW1.pdf" for the document and "your_last_name_HW1_source_code.ipynb" for the source code.
- Before submitting the source code, please double-check that it runs without any errors.
- Must submit the files separately.
- Do not compress into a zip file.
- HW submitted more than 24 hours late will not be accepted for credit.