Homework 5-autoencoder

Single-cell RNA-Seq (scRNA-Seq) data is incredibly noisy and sparse, with only a fraction of transcribed RNAs captured in the sequencing due to the difficulty in amplifying such a small signal from a single cell. Specifically, there are many genes expressed in a cell at any given time, and two entirely different scRNA-Seq readings could come from cells with identical expression profiles.

In this assignment, you will explore utilizing autoencoders to identify clusters of cells from the counts data. These clusters could represent varying expression profiles in the same tissue sample, such as the unique profiles that you might see in different phases of the cell cycle or different cell types.

`counts.npy` is a scRNA-seq gene expression matrix with `5000` cells and `1000` genes. Cell e_{ij} represents the normalized log gene expression value of `j`-th gene in `i`-th sample. `labels.txt` is the corresponding cluster these cells belong to. There are a total of 3 clusters.

1. Train an Autoencoder (composed of fully-connected layers) to learn a low-level gene expression representation of the cells from the scRNA-Seq counts. Use and compare different latent embedding (embedding learned by the bottleneck layer) sizes of 10 and 50. This is the size of the output of the encoder, and the input to the decoder. (3 pts)

   Hint 1: Use MSE loss. This assignment was tested with ~50 epochs. Your network may take more or less to produce some good clusters.

   Hint 2: For AE's, a good rule of thumb is to have the Encoder's architecture be symmetric to the Decoder's. (Ex. Enc=1000->100->10, Dec=10->100->1000). There is no activation function needed behind the input layer, latent embedding layer and the reconstruction (output) layer.

   Hint 3: Normalization is not needed for this particular data.

2. Report and compare the Mean Squared Error (MSE) between the reconstruction and original data for each latent embedding size. How does the size of the latent space affect the reconstruction MSE? (1 pt)

3. Compare and report the plots of the original data with the plots of the reconstructions. How do the PCA & t-SNE plots of the reconstructions compare with those of the original data? Here we only consider latent embedding size of 10. (1 pt)

   Note: Use `labels.txt` as labels when plotting.

4. Compare and report the plots of the original data and the plots of the latent vectors. How do the various embedding sizes change the quality of clustering? Here we only consider latent embedding size of 10. (1 pt)

Note: Use `labels.txt` as labels when plotting.

5. Try to use the following loss to replace MSE when training your AE model. Compare the reconstruction Mean Squared Error (MSE) with Q2 when latent embedding size of 10. Then plot the PCA & t-SNE plots of the reconstructions and latent vectors with original data. What clonclusion can you get? (2 pts)

   Note: There should be 4 plots in total. Two PCAs (reconstruction v.s. origin, latent embedding v.s. origin), two t-SNE (reconstruction v.s. origin, latent embedding v.s. origin).

```python
import tensorflow as tf
def nonzero_mse_loss(y_true, y_pred):
    # Create a mask for non-zero elements
    mask = tf.cast(tf.math.not_equal(y_true, 0), tf.float32)
    nonzero_count = tf.reduce_sum(mask)  # Count the number of non-zero elements
    # Apply the mask to filter out zero elements and Calculate squared difference
    nonzero_squared_diff = tf.square(y_true - y_pred*mask)
    # Calculate the mean of non-zero squared differences
    nonzero_mse = tf.reduce_sum(nonzero_squared_diff) / nonzero_count
    return nonzero_mse
```