# CMPT 880: Deep Learning
# Simon Fraser University
# Spring 2019
# Homework 2

By Xia Hu

This is a classification problem with 3 class labels (labels are 0, 1, 2). The whole dataset including 15,000 images is divided into train, validation and test sets in the ratio of 6:2:2. The instances are 32 by 32 by 3 RGB images. In our datasets each image is represented by a 3072 length vector, the first 1024 entries of the vector contain the red channel, the next 1024 the green, the final 1024 the blue. The image is stored in row-major order, so the first 32 entries are the first row red channel values. The labels of the test set are not provided. Use the train set to train your model. Then, evaluate your model using the validation set.

Download the pickle files, and load train, validation and test sets from the file. We provide two functions to help you load from and save to pickle file. The data files include:

- trainset.pickle: 9000 training images and their labels.
- validset.pickle: 3000 validation images and their labels
- testset.pickle: 3000 test images without labels

This homework contains 2 parts:

1) Train your classification model. You may use Multilayer Perceptron, Convolutional Neural Networks, Deep Belief Networks, or Logistic Regression models. Use the negative log likelihood (multiclass cross-entropy) as the loss. You may use the framework you installed for homework 1 (e.g. Theano, Tensorflow, Keras, Torch, Caffe, or MatConvNet (MATLAB) ). Use your trained model to predict the labels for the test set.
2) Deal with the data unbalanced problem. The number of instances being labeled '0', '1' and '2' are not equal. In each of the train, validation and test sets, 40% instances are with label '0', 40% instances are with label '1', and only 20% instances are with label '2'. The data imbalance problem is important because machine learning models are sensitive to lack of balance. Please find some solutions to deal with this problem.

You should submit:

1) Your source code as a file (for example a .py or .m file). Please do not submit multiple files. Please add at most five commented lines (less than 100 words) to the beginning of your code and describe your architecture and your solution to data unbalanced problem. Also, write the accuracy that you achieved on the validation set.

2) The predicted labels for the test set. Regardless of the framework that you use, this should be a pickle file. If you use MATLAB, you can easily open your .mat file in python and save it as a pickle file. We provide a pickle file namely 'testlabel.pickle' with all labels set to zero. You should open it in python, replace the labels that you predicted for the test set and then save it as a new pickle file.

Here is the grading criteria:

- Technical correctness — 50%: Your model is configured and trained correctly.
- Originality — 20%: The implementations and ideas are from yourself.
- Test performance — 30%.
- Creativity bonus (up to 20%) – novel clever ideas

The function for loading data:

```
def unpickle(file):
    with open(file, 'rb') as fo:
        dict = pickle.load(fo, encoding='bytes')
    data = dict['data']
    labels = dict['label']
    return data, labels
```

This function is to replace the test labels and save to pickle file:

```
def saveTest(your_predict_labels):
    f = open('testlabel.pickle','rb')
    labels = pickle.load(f)
    f.close()
    #here you must replace the labels that you have predicted for the test set:
    for i in range(len(labels)):
        labels[i] = your_predict_labels[i]
    #save the pickle file that you should upload:
    f = open('testlabel.pickle', 'wb')
    pickle.dump(labels, f)
    f.close()
```