ML2 Assignment 2: Reports and Analysis

Task 1(a):

1. Accuracy on Test Data by Varying Optimization Techniques:
    a. Vanilla SGD : 10%
    b. SGD with momentum: 10%
    c. Adam: 48%
2. Accuracy on Test Data by Varying Normalization Techniques:
    a. Dropout: 46%
    b. Batch Normalization: 39%
3. Accuracy on Test Data by Varying Activation function between CNN layers:
    a. Identity: 53%
    b. Sigmoid: 48%
    c. Relu: 55%
    d. TanH: 54%
4. Accuracy on test data by varying Loss Functions:
    a. CEL : 46%
    b. MSE : 8%
    c. L1 : 11%

Best Configuration for the CNN model as per accuracies is Adam optimizer, ReLU Activation, CEL Loss and Batch Normalization.

Explanation: Since we got accuracies high as a strong support for this configuration, I tried out training same model in a single notebook with same configuration and found that Adam optimizer, ReLU Activation, CEL Loss and Batch Normalization results with 55% accuracy. Moreover when I tried same configuration but changing normalization to dropout it gave 54% accuracy.

Lets defined the terms firstly,

1. **Vanilla SGD:** The basic stochastic Gradient Descent Algorithm, the gradient is equal to the l1 norm of gradient at each data point. Random Gradient Descent. SGD randomly picks one data point from the whole data set at each iteration to reduce the computations enormously.
2. **SGD with momentum**: There is a challenge in vanilla SGD that is addressed by SGD with momentum.
    a. Proper Learning Rate: Small learning rate leads to slow convergence, large learning rate will oscillate around minima or diverges.
    b. Learning rate Schedules: changing learning rate according to some predefined schedule.
    c. The same learning rate applies to all parameter updates
    d. The data may be spase ,and different features may have different frequencies.
    e. Larger updates for rarely occurring features updates is a better choice. This is called momentum.

    f. Avoding getting trapped in some suboptimal local optima also called saddle points where in one dimension the slope rises and in some other direction sloe decreases.
    g. SGD with momentum accelerates the coverging power for valinna SGD with such methods.
3. **Adam:** A variant of combination of RMS Prop and momentum incorporates first order moment with exponential weights of gradients. Momentum in RMS prop by adding momentum to rescaled gradients. RMS Prop works on the principle that instead of taking accumulative Sum of Squares of gradients starts from t=0, the RMS Prop takes exponential decaying average of the squared gradient and it also does not consider extreme histories when accumulating small gradients. Once the local convex bowl is found it converges super fast. The decay go on reducing exponentially and that is advantage of taking the exponential decaying average of square gradients.
4. **Dropout:**
    a. Regularization technique
    b. During training randomly selects neurons are dropped from the network with probability 0.5
    c. Their activations are not passed to the downstream neurons in the forward pass
    d. In the backward pass weight updates are ignored for these neurons.
    e. Major drawback is while training the weights of neurons are tuned for specific features that provides some specific specialization. It might miss some features
    f. This lead to specificity to training data.
    g. The dropped neurons penalty is pushed to working neurons.
    h. Network may end up learning different unwanted features.
    i. Avoids Overfitting
5. **Batch Normalization:**
    a. BN addresses the problem of covariate shift, you train the model with batches with disjoint data.
    b. In neural network each class is learned by discriminating the functions for n classes. When dataset of same class is so varied that our decision boundaries hop from one to another. In simple terms for example you have a dataset with flower class. There is high chance of having large number of different types of flowers with varying height, width, color etc. In that case, you end with covariate shift. When you represent all in vector space you get different distribution of flowers itself. Therefore the BN breaks this and says, I will create multiple batches and then put disjoint dataset in all batches, a flower in batch 1 will be quite different in look than flower in batch 2.

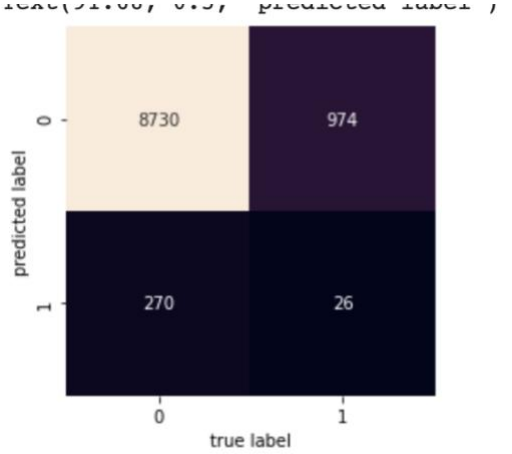
Coming to the question, of best configuration, in our case, we had not used dataset with peculiar and minute changes in within class data. Thus it was expected that dropout shall perform better.

Recollecting everything

1. **Adam performed better because it is the best optimization among all since it have very fast convergence.**

2. **Dropout is meant to block information from certain neurons completely to make sure the neurons do not co-adapt. While BN is used so that we get disjoint sets of examples in training phase. Moreover when I trained with Dropout and BN both resulted approx. same results, 55% and 54% respectively.**
3. **ReLU is the best and popular activation function. This is because it does not activate all neurons at the same time. This means the neurons will only be activated when the output of the linear transformation is greater than 0. F(x)=max(0,x). In our case the property of ReLU that it does not activate all neuron brought higher accuracy. Due to this reason, during the backpropagation process, the weights and biases for some neurons are not updated.**
4. **CEL is preferred for classification while MSE is preferred for regression. We are doing a classification problem in our case hence we got higher accuracy with CEL.**

Task 2(a): `Accuracy: 87.56`




```
Confusion Matrix
[[8730  270]
 [ 974   26]]
Classification report
              precision    recall  f1-score   support

           0       0.90      0.97      0.93      9000
           1       0.09      0.03      0.04      1000

    accuracy                           0.88     10000
   macro avg       0.49      0.50      0.49     10000
weighted avg       0.82      0.88      0.84     10000


Confusion Matrix:
[[8730.  974.]
 [ 270.   26.]]
Accuracy: 87.56
```