DD2424 DEEP LEARNING IN DATA SCIENCE
Assignment 2

Rithika Harish Kumar(rihk@kth.se)

AIM

To train and test a two-layer network on a multi classification problem with stochastic mini batch gradient descent. Momentum is also applied to fasten the training while the aim is to minimize the loss function called cross entropy with a regularization term. The dataset to be used is CIFAR-10.

RESULTS

Gradient check:

Numerical gradients were computed by computing a base value for the cost function, making a small variation to a single weight value, recomputing the cost, estimating the derivate and resetting it. In this manner it is calculated for every weight in the network. The analytic gradients and the numerically computed gradients are compared by summing the difference between the respective values in the matrix. If the sum value is low it is better. Convinced that the implementation is correct because by choosing few hyperparameters the loss function converges to a satisfactory level.

Two-layer network with 0.0 and 0.0 regularization

```
Weight1
Difference: 2.710e-04
Mean of the values   Analytical: 6.622e-03   Numerical: 6.622e-03
Min of values   Analytical: 0.000e+00   Numerical: 0.000e+00
Max of values   Analytical: 5.580e-02   Numerical: 5.580e-02

bias1
Difference: 2.997e-07
Mean of the values   Analytical: 1.219e-02   Numerical: 1.219e-02
Min of values   Analytical: 0.000e+00   Numerical: 0.000e+00
Max of values   Analytical: 3.077e-02   Numerical: 3.077e-02

Weight2
Difference: 9.677e-07
Mean of the values   Analytical: 9.121e-03   Numerical: 9.121e-03
Min of values   Analytical: 0.000e+00   Numerical: 0.000e+00
Max of values   Analytical: 1.234e-01   Numerical: 1.234e-01

bias2
Difference: 4.472e-07
Mean of the values   Analytical: 3.648e-02   Numerical: 3.648e-02
Min of values   Analytical: 4.552e-03   Numerical: 4.552e-03
Max of values   Analytical: 9.601e-02   Numerical: 9.601e-02
```

Two-layer network with 0.1 and 0.3 regularization

```
Weight1
Difference: 4.775e-03
Mean of the values   Analytical: 1.873e-02   Numerical: 1.873e-02
Min of values   Analytical: 8.785e-08   Numerical: 1.421e-08
Max of values   Analytical: 1.253e-01   Numerical: 1.253e-01
```

DD2424 DEEP LEARNING IN DATA SCIENCE
Assignment 2

Rithika Harish Kumar(rihk@kth.se)

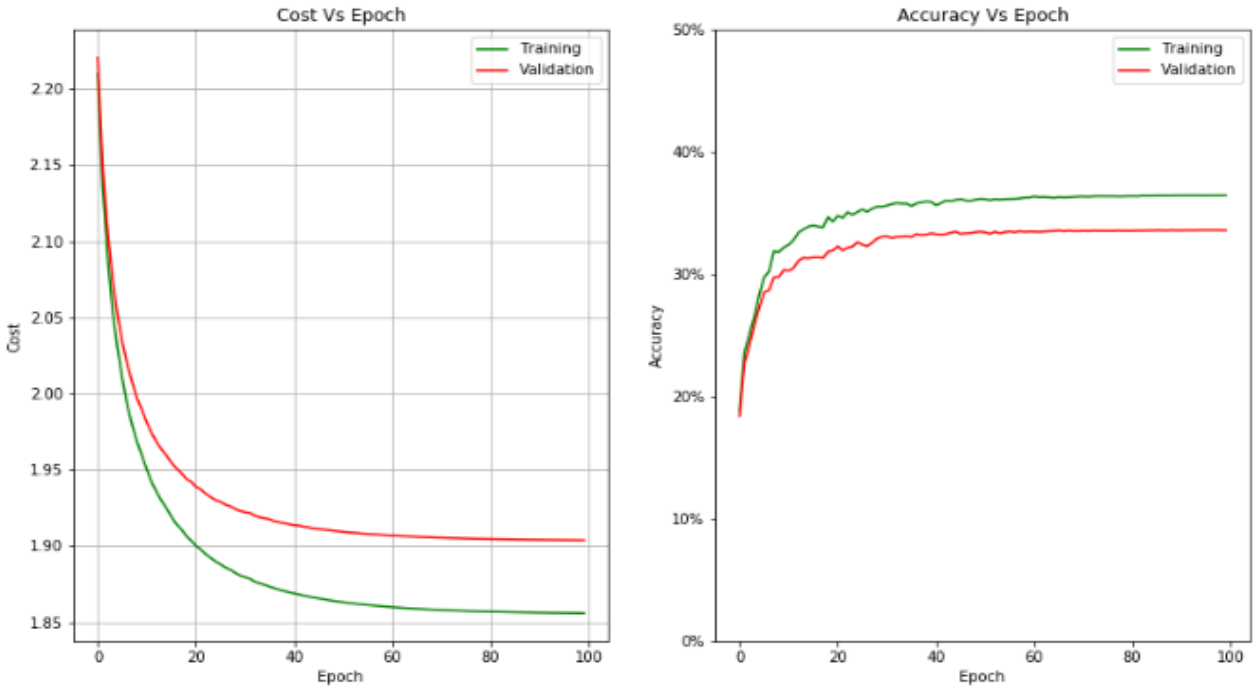
```
bias1
Difference: 2.071e-07
Mean of the values   Analytical: 3.517e-02   Numerical: 3.517e-02
Min of values   Analytical: 0.000e+00   Numerical: 0.000e+00
Max of values   Analytical: 1.741e-01   Numerical: 1.741e-01

Weight2
Difference: 4.590e-05
Mean of the values   Analytical: 1.482e-01   Numerical: 1.482e-01
Min of values   Analytical: 5.435e-03   Numerical: 5.436e-03
Max of values   Analytical: 4.607e-01   Numerical: 4.607e-01

bias2
Difference: 4.441e-07
Mean of the values   Analytical: 5.292e-02   Numerical: 5.292e-02
Min of values   Analytical: 3.601e-03   Numerical: 3.601e-03
Max of values   Analytical: 1.198e-01   Numerical: 1.198e-01
```

Effect of momentum:

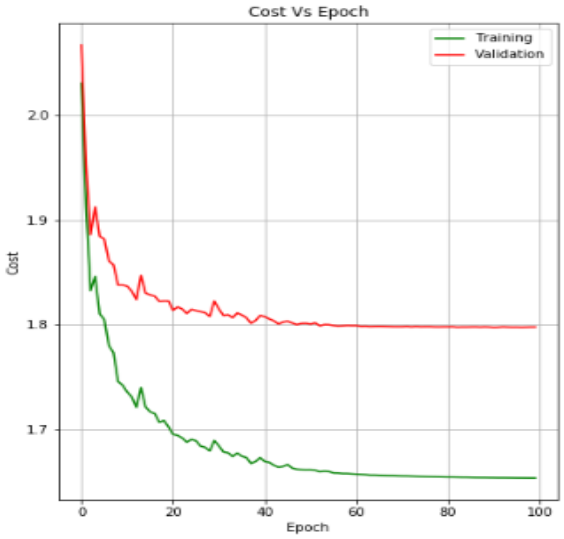
Without momentum: eta_initial=0.01, eta_decay=0.95

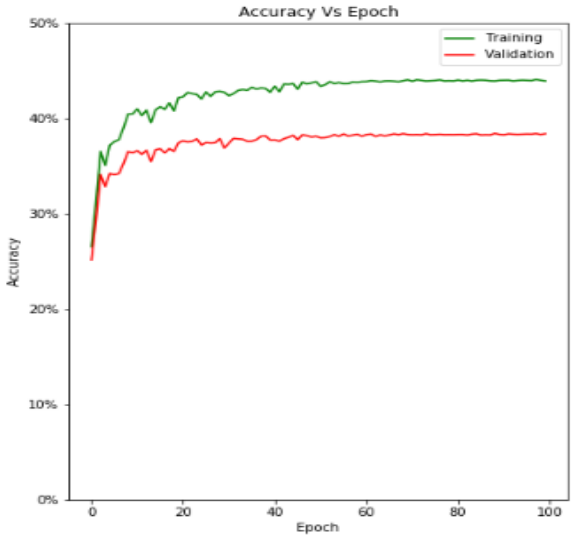


After Adding Momentum: eta_initial=0.01, eta_decay=0.95, rho = 0.9

DD2424 DEEP LEARNING IN DATA SCIENCE
Assignment 2

Rithika Harish Kumar(rihk@kth.se)



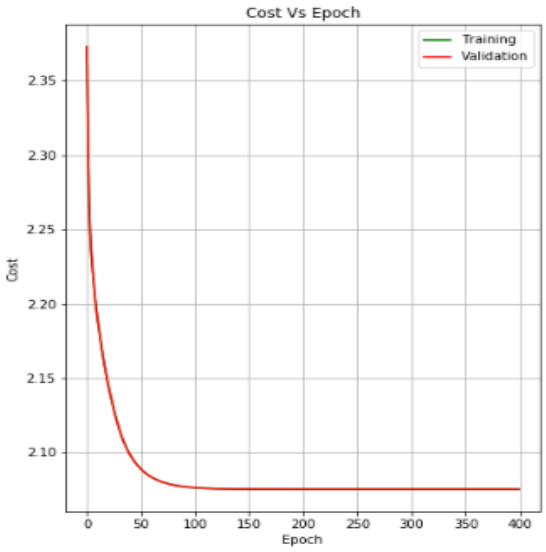


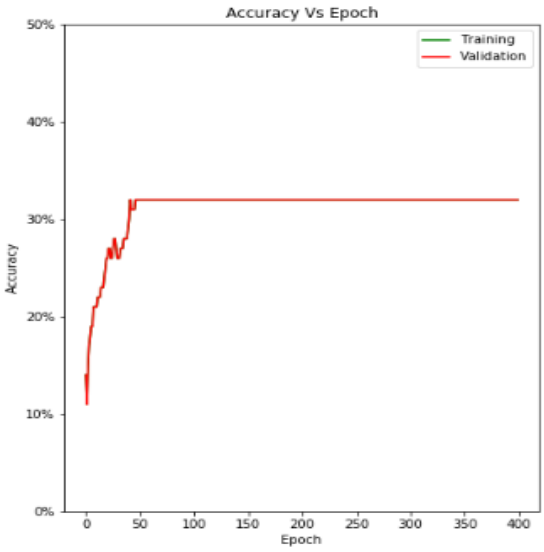
Vanilla SGD without momentum is slow compared to the momentum update. It updates the network parameters not only using current gradient but also with weighted history of the previous updates. That weight of past history is called momentum.

With overfitting:

By taking a small subset of data and training for high number of epochs overfitting is done and tested for different rho values with eta_initial=0.005, eta_decay=0.95. Increase in momentum increases the speed of convergence. But a very high momentum leads to gradient oscillation.

Rho = 0.5

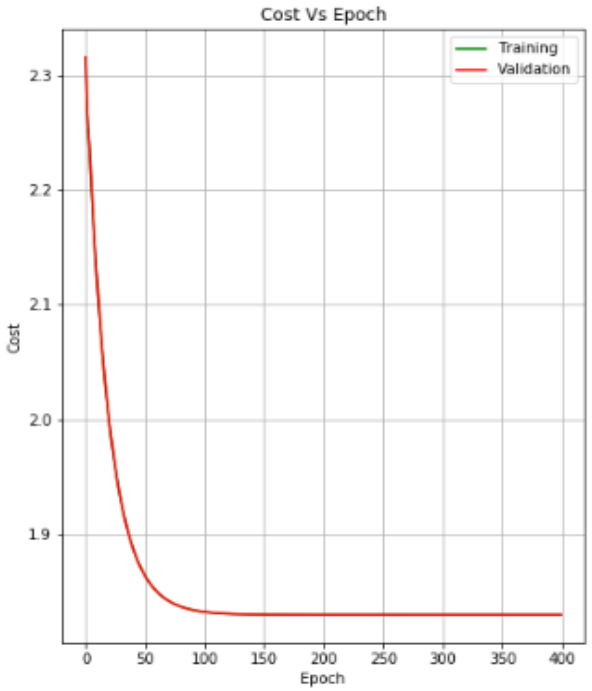


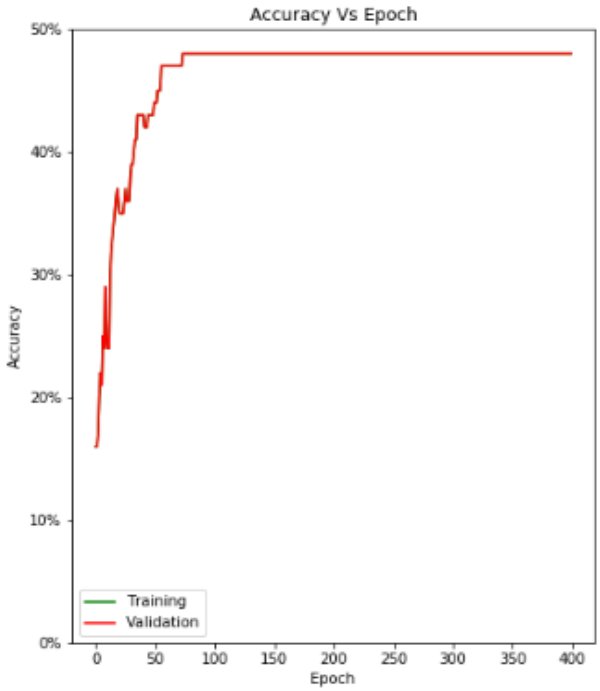


Rho = 0.8

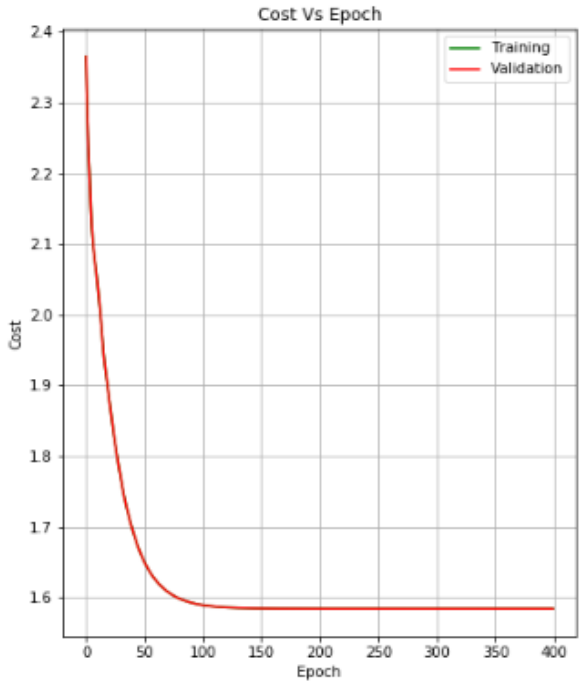
DD2424 DEEP LEARNING IN DATA SCIENCE
Assignment 2

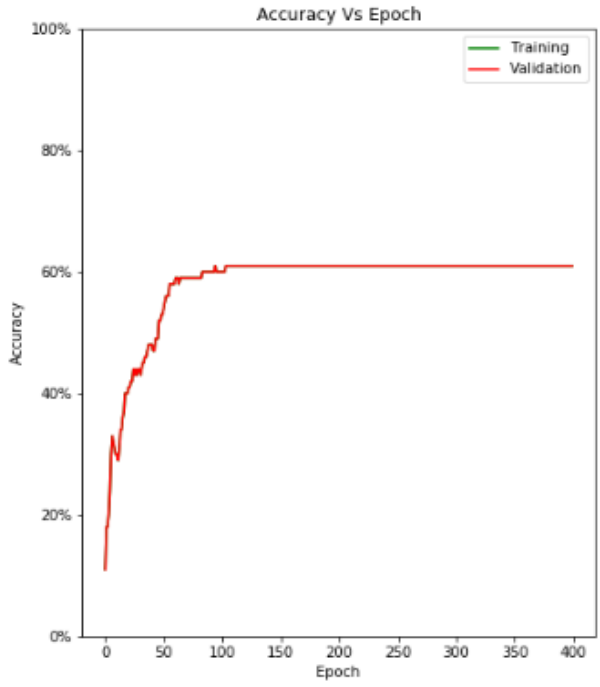
Rithika Harish Kumar(rihk@kth.se)





Rho = 0.9

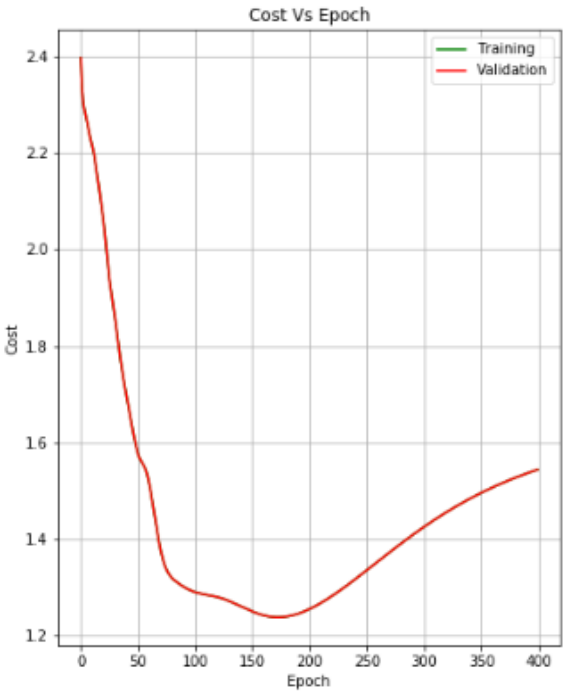


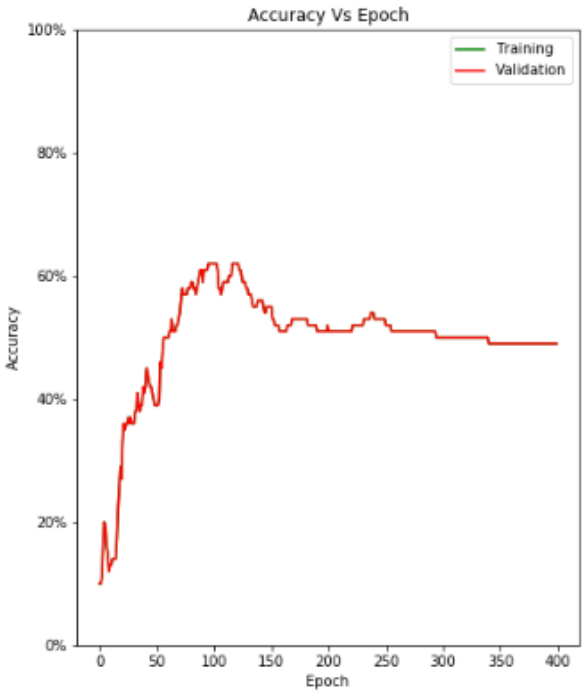


DD2424 DEEP LEARNING IN DATA SCIENCE
Assignment 2

Rithika Harish Kumar(rihk@kth.se)

Rho = 0.99





Best parameter search:

Coarse search

The basic hyper-parameters are set as:

rho = 0.9
eta_decay = 0.95
hidden_size = 50
epochs = 5

The regularization and initial learning rate have 49 combinations:

lamda = [0.001, 0.005, 0.01, 0.05, 0.1, 0.3, 0.5]

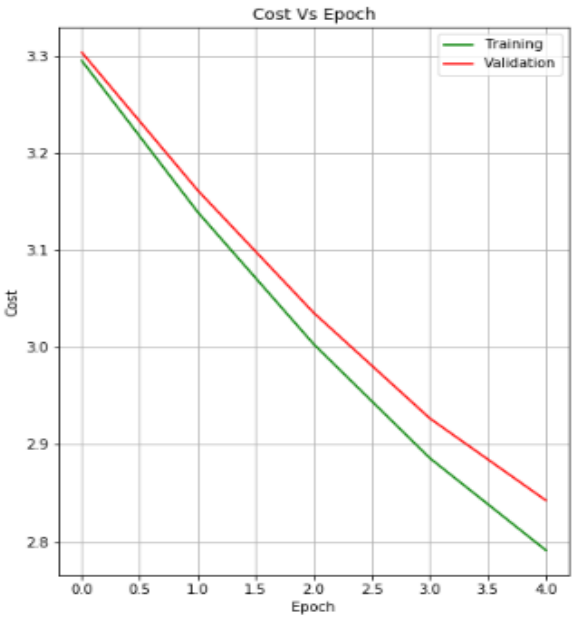
eta_initial = [0.5, 0.3, 0.1, 0.05, 0.01, 0.005, 0.001]

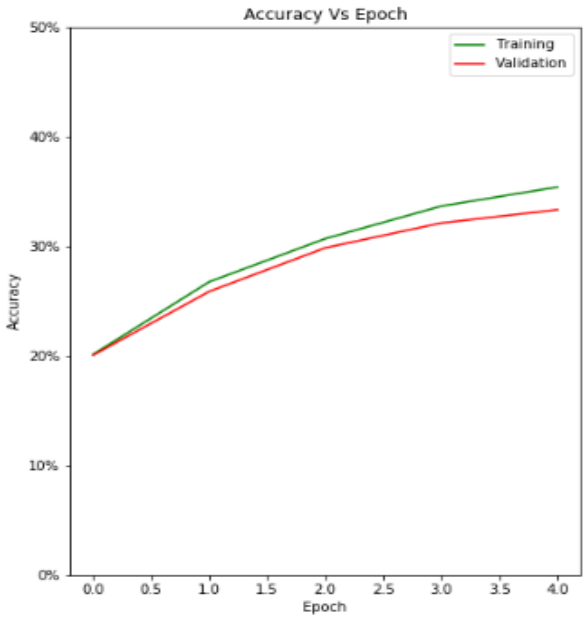
Better performing three networks are:

- eta_initial = 0.5
  lamda = 0.01
  Training accuracy = 35.42%
  Validation accuracy = 33.35%

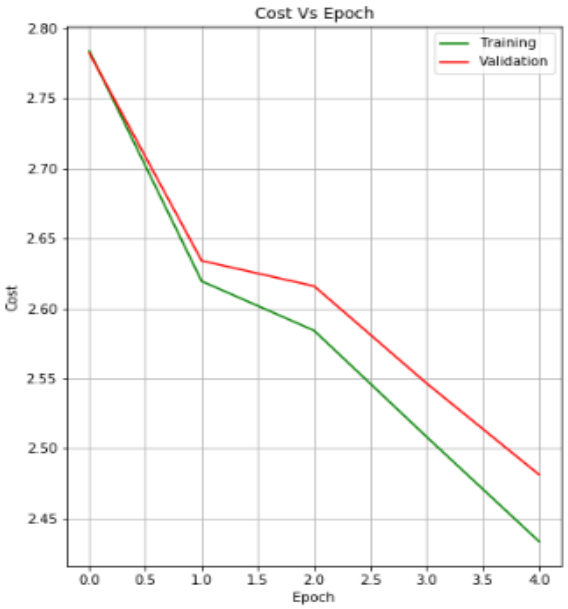
DD2424 DEEP LEARNING IN DATA SCIENCE
Assignment 2

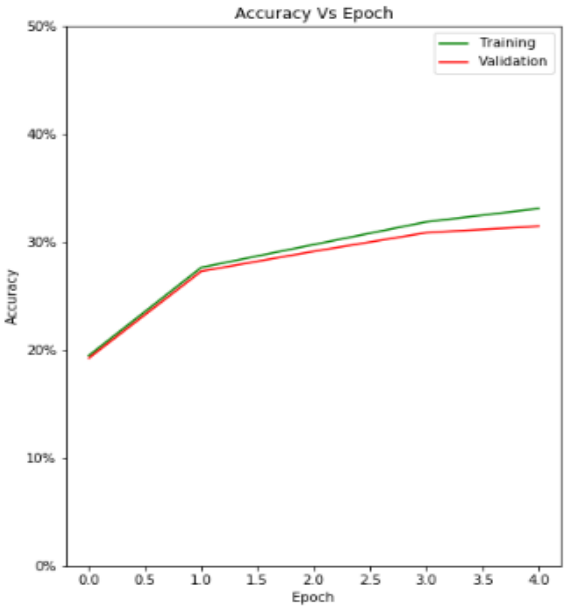
Rithika Harish Kumar(rihk@kth.se)





- eta_initial = 0.5
  lamda = 0.005
  Training accuracy = 33.11%
  Validation accuracy = 31.47%

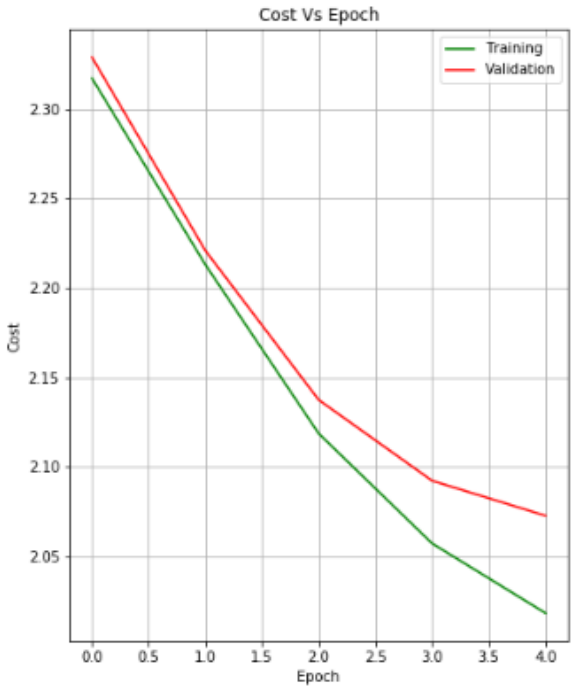


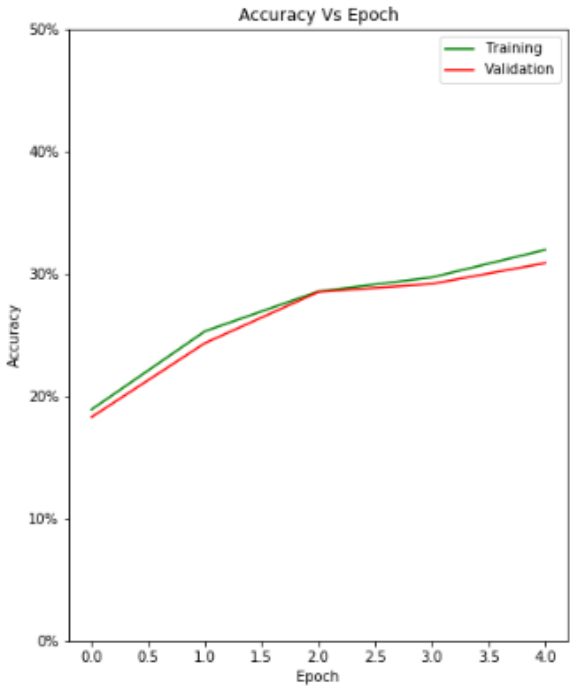


- eta_initial = 0.5
  lamda = 0.001
  Training accuracy = 31.97%
  Validation accuracy = 30.89%

DD2424 DEEP LEARNING IN DATA SCIENCE
Assignment 2

Rithika Harish Kumar(rihk@kth.se)





Fine search

The basic hyper-parameters are set as:

Rho = .9
eta_decay = 0.95
hidden_size = 50
epochs = 10

The regularization and initial learning rate have 28 combinations:

lamda= np.linspace(0.005, 0.0005, num=7)

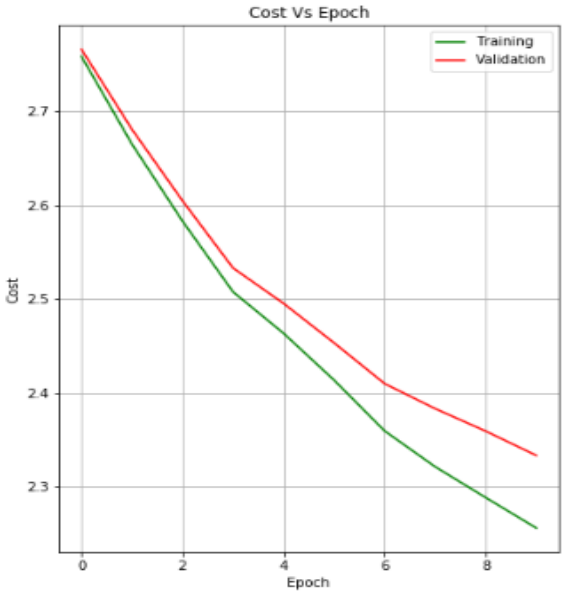
eta_initial = np.linspace(0.35, 0.25, num=4)

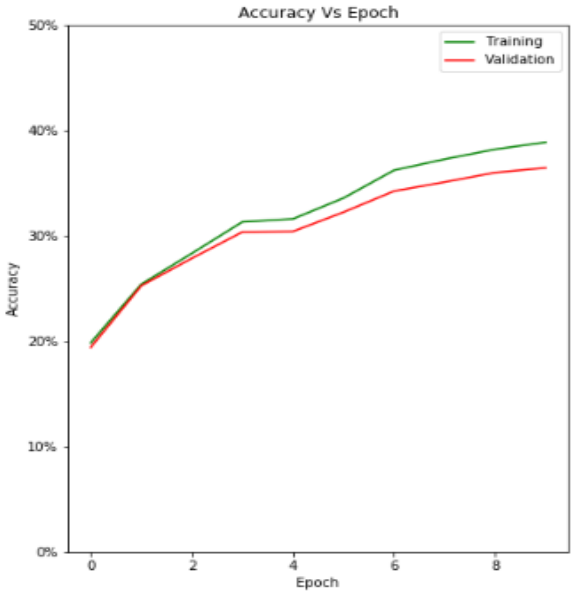
Better performing three networks are:

- eta_initial = 0.35
  lamda = 0.005
  Training accuracy = 38.89%
  Validation accuracy = 36.46%

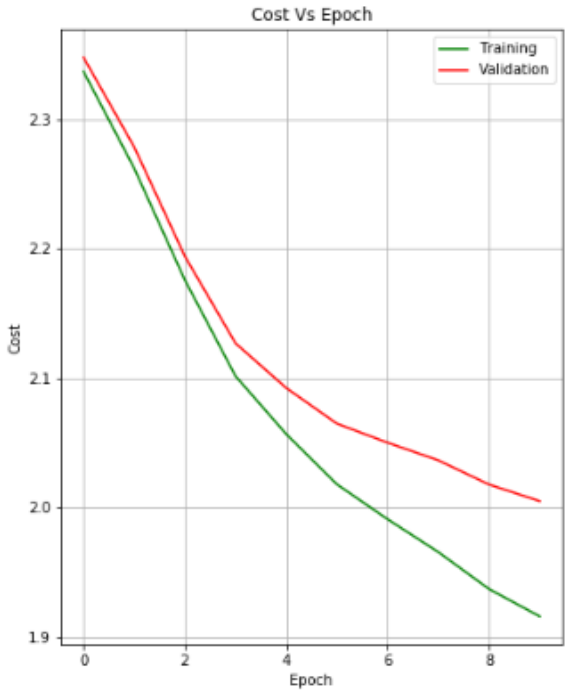
DD2424 DEEP LEARNING IN DATA SCIENCE
Assignment 2

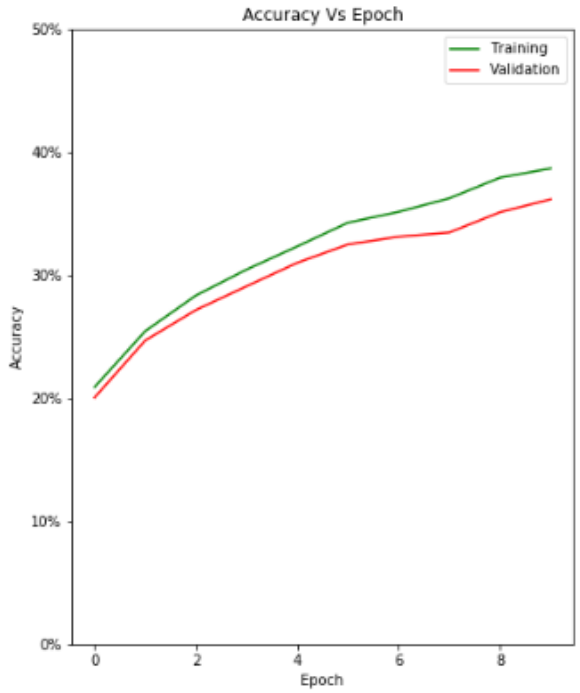
Rithika Harish Kumar(rihk@kth.se)





- eta_initial = 0.35
  lamda = 0.00125
  Training accuracy = 38.69%
  Validation accuracy = 36.18%



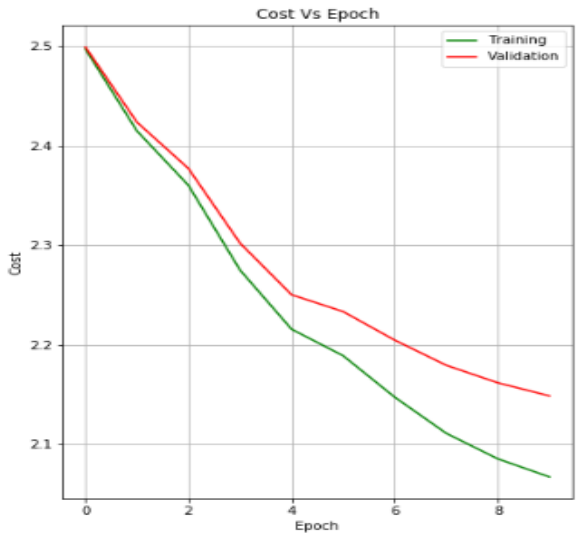


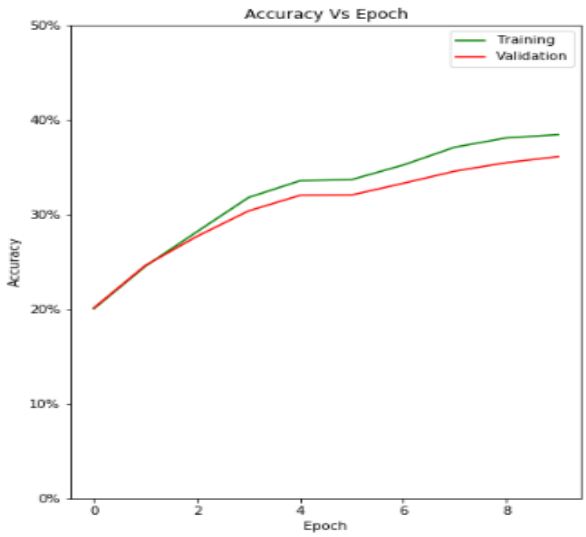
- eta_initial = 0.35

DD2424 DEEP LEARNING IN DATA SCIENCE
Assignment 2

Rithika Harish Kumar(rihk@kth.se)

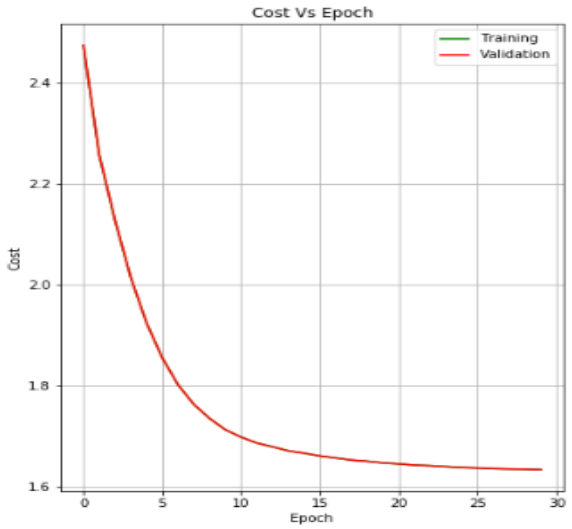
lamda = 0.00275
Training accuracy = 38.45%
Validation accuracy = 36.13%

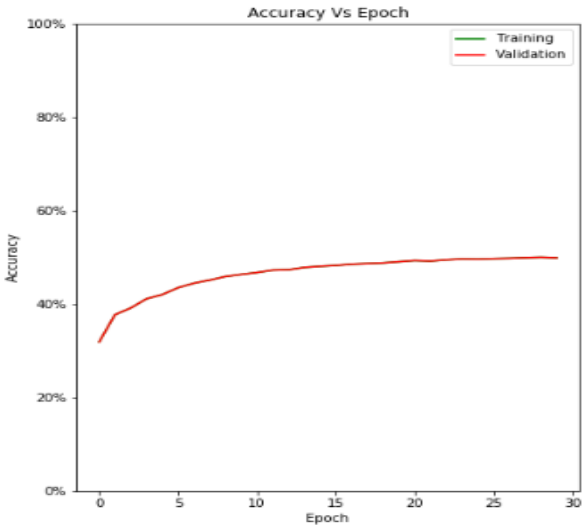




The best network according to the highest validation accuracy obtained is trained where the whole dataset is used for training, leaving out the final 10000 images for testing.

Epochs = 30
lamda=0.005
eta_initial=0.35
eta_decay=0.95 and rho=.9.





Training accuracy = 49.90 % **Testing accuracy = 48.53 %**