Assignment 2

Albin Larsson Forsberg

April 16, 2020

1 Introduction

The assignment consists of creating a basic one layer network that is classifying images from the CIFAR-10 dataset. The inputs consists of a input layer of size 3072, a hidden layer of size 50, and an output layer of size 10 with activation function being a softmax. Whichever node has the highest probability assigned to it will be the prediction of the network.

2 Dataset

The dataset used is the CIFAR-10 dataset that is coming from the Canadian Institute For Advanced Research. It contains in total 60,000 picture from 10 categories: airplanes, cars, birds, cats, deer, dogs, frogs, horses, ships, and trucks. It is widely used in Machine Learning to develop and train different classification models.

3 Model

The Model used is a simple model with only three layers, an input, a hidden, and an output layer. The input consists of the rgb data from each pixel and the three dimensional Matrix structure has been vectorized to only be in one dimension. This vector is the fed into the network that predicts the classifier. The output layer is defined by function 1, while the hidden layer is defined by function 2. Where W are the weights and b the biases. The cost is defined by the cross entropy function.

$$f(X) = \text{softmax}(W^{[2]}a^{[1]} + b^{[2]}) \tag{1}$$

$$f(X) = \text{ReLU}(W^{[1]}X + b^{[1]}) \tag{2}$$

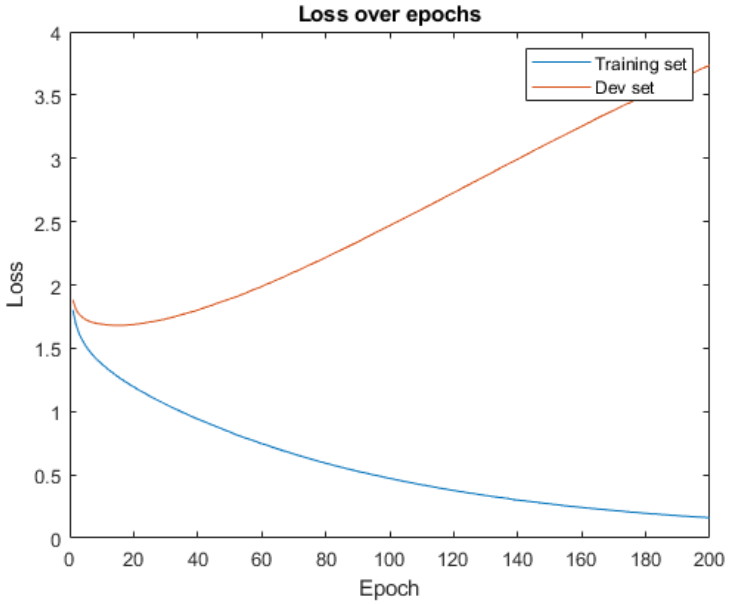

Figure 1: Example of overfit when training for too long

4 Results

The gradient was calculated analytically by hand and was implemented in code. An indication that the gradient is correct is from the way that the data overfits to the training data as can be seen in figure 1. Actually comparing the gradients numerically gives a relative error between the gradients in the scale of 10^{-7}, almost 10^{-8}, which is seen as good enough for the purposes of this problem. Worth noting is that the relative error is larger for the weights that are earlier in the network compared to the ones that are towards the end.

The step size is a cycling version that alternates between two values, η_{min} and η_{max}. The cyclic pattern of η is seen in figure 2. Using this cyclic η lead to the result seen in figure 3, and 4. What can be seen in these graphs is that during the part of the curve where the step size gets larger, the training is accelerating, but after one cycle, we get a large error due to the larger step size. When it afterwards goes down again we usually end up in a better spot. An advantage with this cyclic pattern is that the model during training, can get out of a local optima if needed and jump to find a better one. That is the behavior that we see happening. The accuracy after one period on test data is 0.5230, and after the full three periods 0.5280

For the coarse search values between 10^{-5}, and 10^{1}, were used. We can see that too much regularization means that the model is no better than a pure guess. The three most optimal values of lambda can be seen in table 1. A plot of accuracy vs λ is seen in figure 5. 50 different values of lambda were simulated

Lambda	Accuracy
$2.1 \cdot 10^{-4}$	0.485
$1.9 \cdot 10^{-4}$	0.482
$6.1 \cdot 10^{-5}$	0.480

Table 1: Result from coarse Lambda search

Lambda	Accuracy
$1.28 \cdot 10^{-4}$	0.495
$4.60 \cdot 10^{-4}$	0.495
$2.89 \cdot 10^{-4}$	0.493

Table 2: Result from fine Lambda search

on a training set of 45000 images that were randomly drawn from all of the 50000 training samples. The coarse search was training for 2 cycles.

For the fine search 20 values between 10^{-5}, and 10^{-3}, were used. The three most optimal values of lambda can be seen in table 2. A plot of accuracy vs λ is seen in figure 6. The optimal value of lambda found was $\lambda = 1.283 \cdot 10^{-4}$. The fine search was training for 2 cycles.

The best performing lambda created a model that was trained and provided the curves seen in figure 7. The network performed with an accuracy of 48.63%on the test set. The accuracy on devset and training set are 51.10% 53.62% respectively.

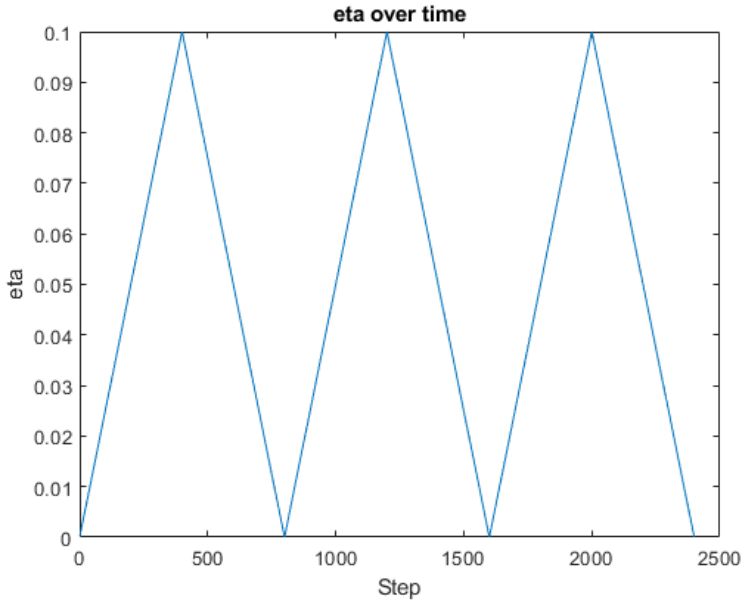

Figure 2: Periodic pattern of η

.

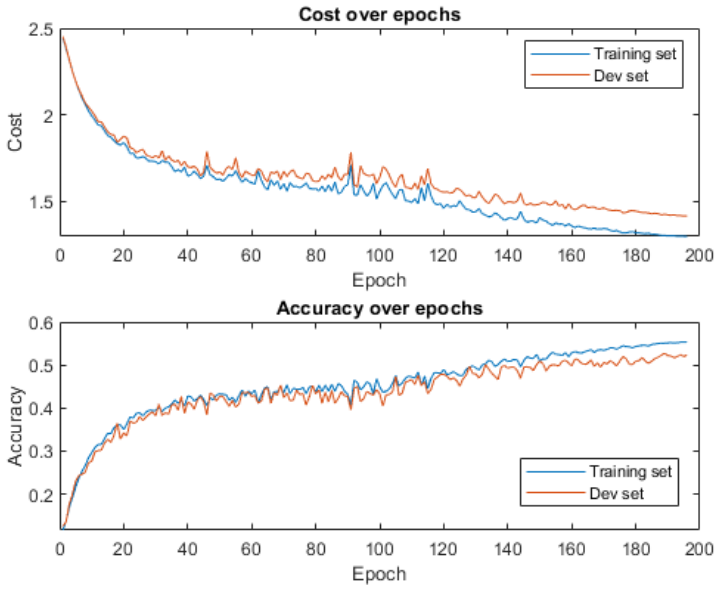

Figure 3: Cost and Accuracy over one period of training.

.

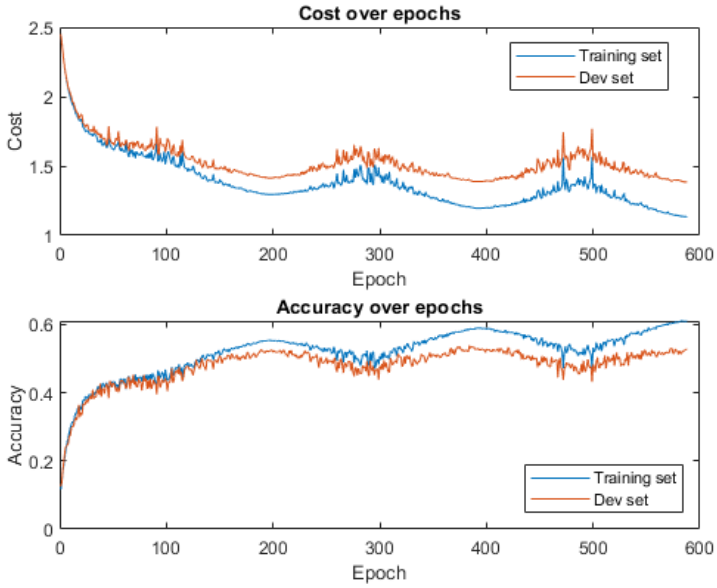

Figure 4: Cost and Accuracy over three periods of training.

.

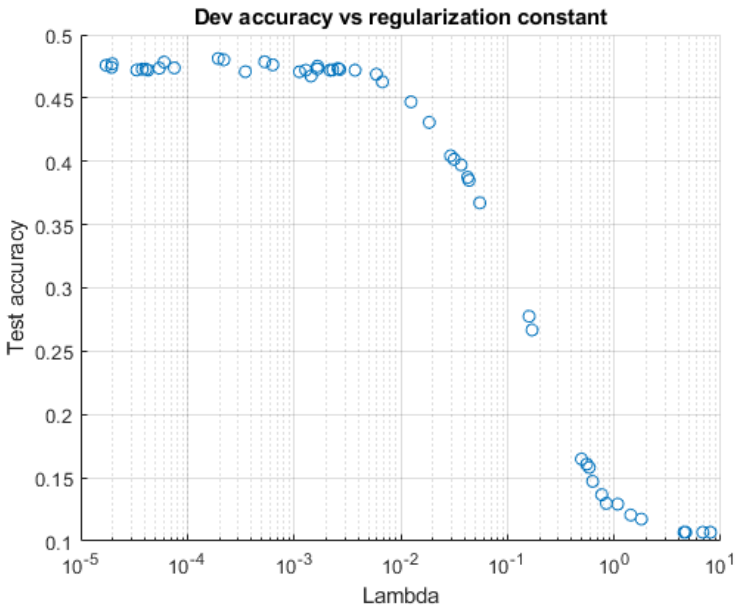

Figure 5: Validation accuracy versus different values of regularization for a coarse search.

.

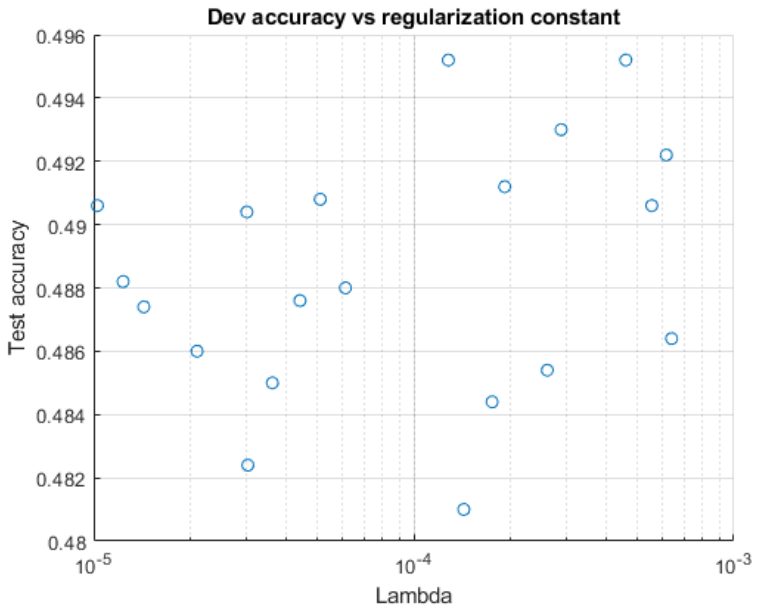

Figure 6: Validation accuracy versus different values of regularization for a fine search.

.

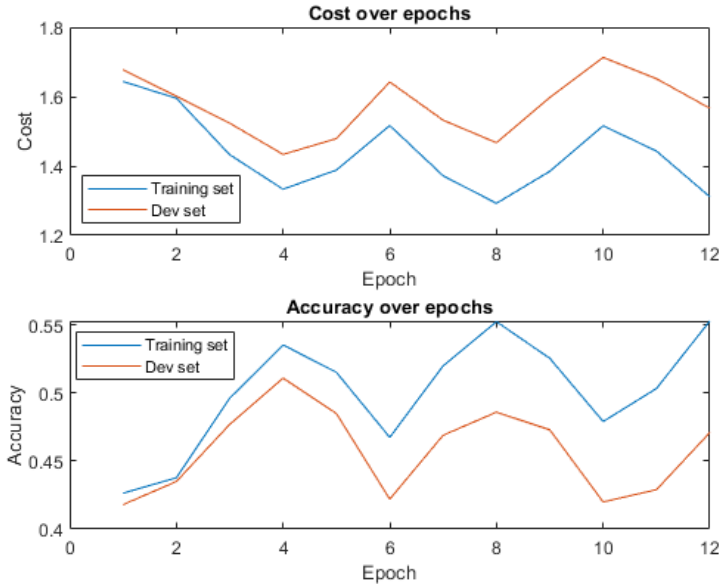

Figure 7: Validation and training Cost for good lambda.

.