An Introduction to Neural Networks for Classification of the Fashion-MNIST Dataset

Jeffrey M Abraham

ABSTRACT

The purpose of this report is to design fully-connected and convolutional neural networks to classify images of clothing into their types.

Keywords: Tensorflow, SGD, MNIST,

I INTRODUCTION AND OVERVIEW

This analysis is being done to classify 28x28 pixel images of articles of clothing into one of ten categories. The MNIST Fashion data set is meant to be a benchmark test for machine learning algorithms, and is mo

II THEORETICAL BACKGROUND

Fully Connected Neural Networks

A fully connected neural network as pictured in Figure 1 works by multiplying each input by a matrix of weights that determines the value of each neuron in the following layer. In this case each neuron in the hidden layer is determined by four weights multiplied by each of the inputs. Not pictured are bias neurons that add constants to the weighted inputs to shift the outputs. This has a similar function to the b term in the equation for a line $y = mx + b$. Activation functions are applied after each layer. The output layer has a specific activation function called the *softmax* activation function. This gives outputs as probability values. The neural network first has to be trained which is done by using some variation of gradient descent to minimize a loss function. For classification purposes, the neural network shown in Figure 1 would be trained using a number of (X,y), with four values in each X and two values in each y. The loss function simply put takes the average predicted probability that each input is its corresponding output in vector y. Gradient descent is done to find the way that changing each weight value or bias value would change the loss function.

$$\vec{y} = A\vec{X} + \vec{b} \tag{1}$$

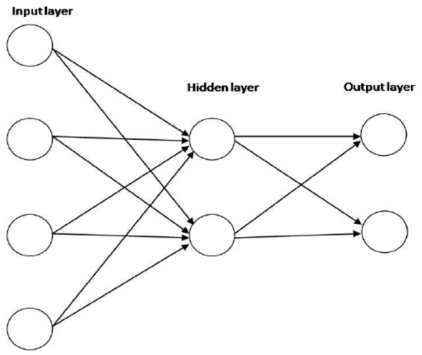


Figure 1. A Fully Connected Neural Network with 1 Hidden Layer

Optimization Functions

There are many varieties of optimization functions used to minimize the loss function. Most of these, however, are just variations of Gradient Descent. The purpose of an optimization function is to successively get closer to some minimum loss. Many hyper-parameters can be used to fine tune these functions including but not limited to learning rates, momentum, and initialization schemes.

Stochastic Gradient Descent

Stochastic gradient descent is a variation of gradient descent that uses a randomly selected data point to calculate gradients rather than the entire data set. This can cut down on processing power, but also can take indirect paths to local minima. Some compromise can be made by splitting data randomly into small batches to be used for training. This is still less intensive on the computer being used to train the network, but also allows for more direct paths to minima and less variation after arriving at a minimum.

Learning Rate

The learning rate sets the magnitude that the weight and bias values move in the direction set by the gradient. Too high learning rates can cause loss functions to skyrocket, or fall quickly and oscillate around a minima. Using something called learning rate scheduling can change the learning rate between iterations to speed up learning, while not compromising on accuracy.

Back Propagation

Back Propagation is a Jargon term used to describe the chain rule used for calculating the gradient of the loss function. It starts at the output and calculates partial derivatives and moves backward through the network to calculate partial derivatives using chain rule and the previously calculated partial derivatives.

Activation Functions

Activation functions are used to shape the output of each layer of a neural network to make them more friendly as inputs of the following layer, or in the case of output activation functions to transform values into probabilities.

Cross Validation

Cross validation is a method for checking accuracy of a predictive model. It requires that you split your data randomly into training and test data. Standard training:test splits are 70:30 and 80:20. Then you can run the test data through the predictive model and see how accurate your model is. This is important because it is possible to fit a model so that it works perfectly for the training data, but as soon as you test your model you get bad results. This is called over-fitting, and cross validation is an effective method for testing the fit of a model. For neural networks, using validation data and test data can further prevent over fitting by allowing you to tune hyper parameters with validation data and check model accuracy with test data. This way if you over-fit your model to the validation data it will still preform poorly on the test data.

Convolutional Layers

Convolutional layers in a nutshell sweep across an input signal and apply filters to areas of the signal to look for features. In a two dimensional context the filters consists of matrices of values between zero and one that exemplify features of interest. In the context of images this can be very useful by identifying features regardless of location within an image. If the training data has all the identifying features in a certain area of the image, a fully connected layer relies on pixel values which would change drastically, but with convolutional layers, these features would still be identified and images classified correctly.

III ALGORITHM IMPLEMENTATION AND DEVELOPMENT

Fully Connected Neural Network

I started development by collapsing the input into one dimension and using a pyramid like network architecture with ReLu activation, the Adam optimizer L2 regularization and a fixed training rate. On this baseline model I was able to achieve validation accuracy of around 89%

Network Architecture

I started with two hidden layers, the first 300 neurons wide and the second 100 neurons wide. the input is 28x28 so when flattened is 784 neurons, so this structure progressively reduces the number of neurons that are used to define each input to gradually move towards a reliable classification. To make the network deeper I experimented with having multiple layers at the same size before moving to a narrower layer. Also adding more different width layers like a 50 neuron wide layer right before the 10 wide output layer.

Activation Functions

Some experimentation was done exchanging the ReLu activation function which just returns zero for all negative output values to something like ELu the exponential linear unit which uses an exponential function on all negative valued tensors. This did not result in better results and rather made for larger variation in accuracy around the minimum achieved.

Optimization

Some experimentation was done with the Stochastic Gradient Descent activation function but I found that this function took too long to normalize for the fully connected layers. Adam optimizer in combination with learning rate scheduling using a constant learning rate for the first 10 epochs and exponentially decaying after that. This allowed for a sped up convergence without sacrificing stability and accuracy.

Adding Convolutional Layers

To try and improve performance further above 90% accuracy, convolutional layers were used first to identify the features within each image and then fully connected layers were used to train and classify based on the features extracted. I designed my model loosely based on the AlexNet architecture. This model was built to classify color images with much higher resolution so it had to be adapted and shortened to work on the data I was using and run faster given the computer power I am working with.

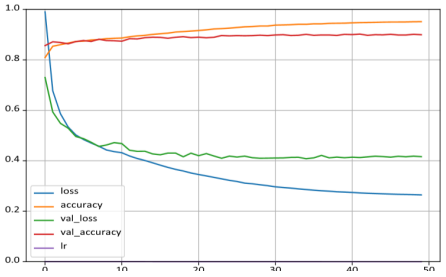
IV COMPUTATIONAL RESULTS

Fully Connected Results

Using only fully connected layers I was able to get results with a maximum validation accuracy of 90.1% using the following hyper-parameters. Two hidden layers of 300 neurons followed by two hidden layers with 100 neurons, both using PReLu activation and L2 regularization with a constant of 0.0008 to prevent over fitting. Scheduled learning rate using 0.0005 for the first 10 epochs and decaying after. After evaluating the test data, an accuracy of 89.3%. The confusion matrix and learning curves can be seen in Figure 2. The confusion matrix shows that the 0 and 6 items are the most often mistaken for each other. This is understandable as these are the T-Shirt/Top and Shirt categories respectively.

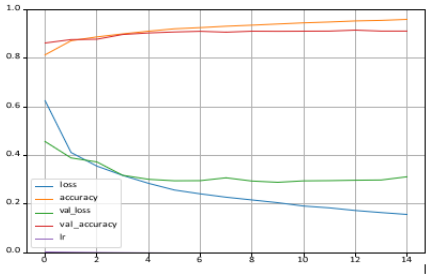
CNN Results

I applied some of the same methods to the convolutional neural network as I did to the fully connected layers. I used scheduling learning rate with the exponential decay beginning at the fifth epoch. Specifics of the network architecture can be seen in the code in the appendix. The best validation accuracy achieved was 91.4% and the test accuracy was 91.2%. The confusion matrix and learning curves can be seen in Figure 3.



	0	1	2	3	4	5	6	7	8	9
0	832	3	19	16	4	2	117	0	7	0
1	2	975	0	14	4	0	4	0	0	1
2	17	0	817	13	82	1	70	0	0	0
3	23	7	10	895	30	2	31	0	2	0
4	1	0	86	32	827	0	53	0	1	0
5	0	0	0	1	0	960	0	21	0	18
6	103	1	82	33	60	0	714	0	7	0
7	0	0	0	0	0	13	0	958	0	29
8	5	2	5	7	4	1	4	8	964	0
9	1	0	0	0	0	7	0	32	0	960

Figure 2. Learning Curves(Left) and Test Confusion Matrix(Right)



	0	1	2	3	4	5	6	7	8	9
0	900	0	18	15	3	2	57	0	5	0
1	0	976	4	12	3	0	3	0	2	0
2	19	0	888	9	44	0	38	0	2	0
3	17	2	12	917	29	0	22	0	1	0
4	4	0	53	23	881	0	38	0	1	0
5	0	0	0	1	0	971	0	15	1	12
6	142	0	75	28	73	0	671	0	11	0
7	0	0	0	0	0	7	0	974	0	19
8	2	0	1	6	6	2	1	4	977	1
9	0	0	0	0	0	4	0	30	1	965

Figure 3. Learning Curves(Left) and Test Confusion Matrix(Right)

V SUMMARY AND CONCLUSIONS

As an introduction to neural networks and hyper-parameter tuning, the limitless options can be overwhelming, but with a basic understanding of what each parameter controls neural network optimizations could be done. With more exploratory time I'm sure that better results could be achieved. The results achieved through this exercises were by no means revolutionary, but I was able to make progress and improve my results.

APPENDIX