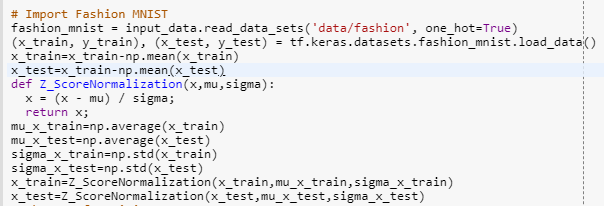
8. (This homework is due: Nov. 15) Experiment convolutional neural network (CNN) for Fashion MNIST data set 1 and generate a report.

(a) Perform “Mean Subtraction” and “Normalization” on the input data. Explain how you perform the two in the report.



Mean Subtraction :Calculate the average of all the examples，then ever example substract the mean.

Normalization: This method gives the raw data mean and standard deviation data standardization. The processed data conforms to the standard normal distribution, that is, the mean value is 0, and the standard deviation is 1.

The conversion function is:



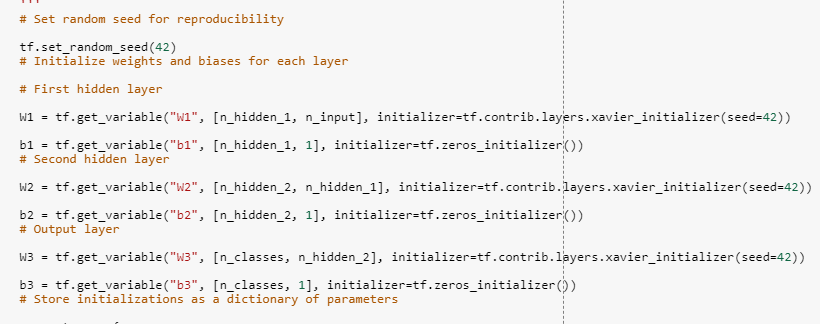
The average can be obtained by the function of np.average(),the sigma can be obtained by the tunciton of np.std().

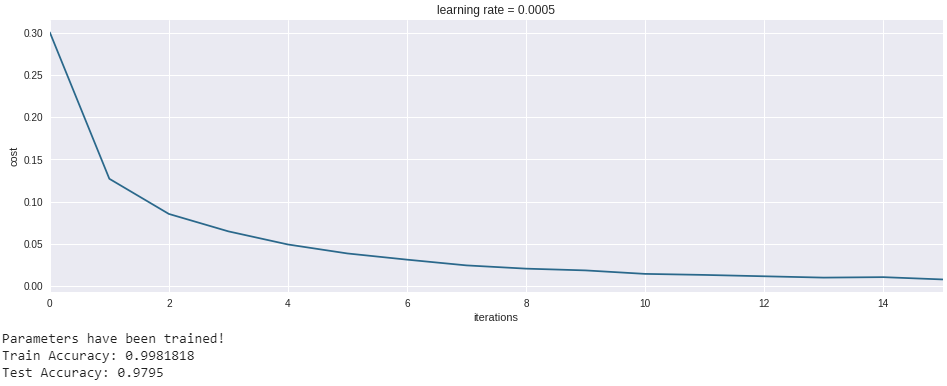
(b) For the initialization, run Xavier initialization and He initialization for the image dataset. Explain how you perform the two in the report with the result table. Show the comparison results in a table. Discuss why you have the result.

i. Xavier initialization appears in a paper named “Understanding the diﬃculty of training deep feedforward neural networks”.

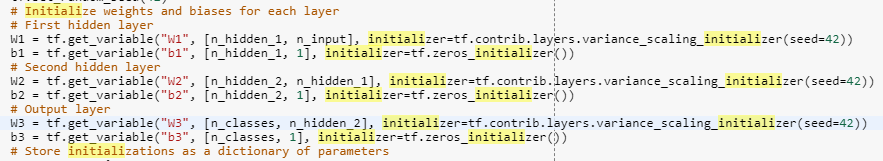
ii. He initialization appears in a paper named “Delving Deep into Rectiﬁers: Surpassing Human-Level Performance on ImageNet Classiﬁcation”.

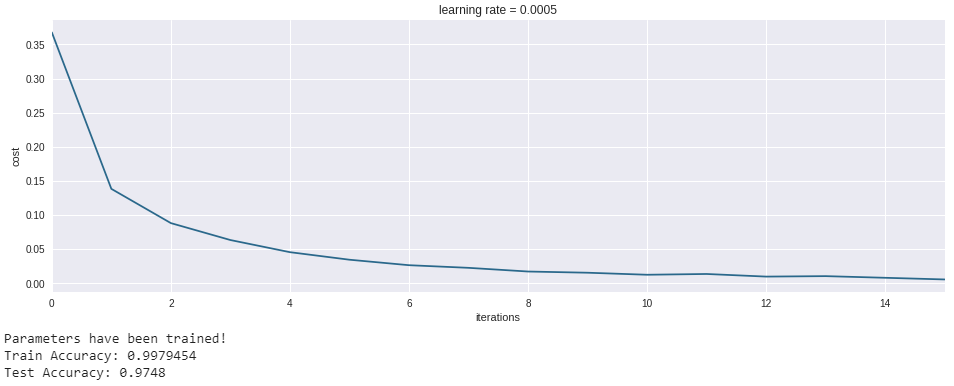
Xavier initialization





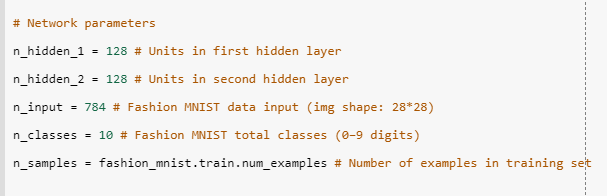
He initialization:

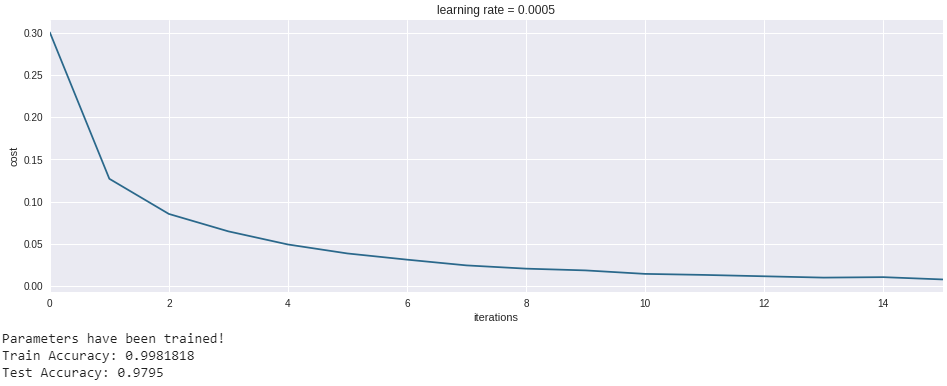


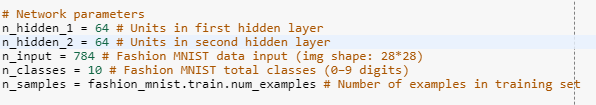


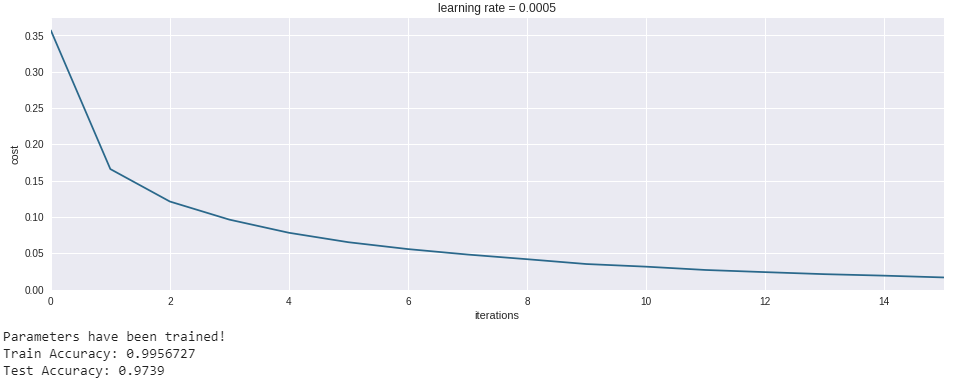
The two results are similar Xavier initialization and He initialization are effective for input data initialization. That’s because the two initialization methods help to make the consistent of input data and output data of each layer.The two results are similar

(c) For the network conﬁguration (how many layers and how many hidden nodes), experiment on at least two diﬀerent conﬁgurations. Explain how you perform the two in the report with the result table. Show the comparison results in a table. Discuss why you have the result.





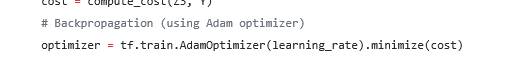


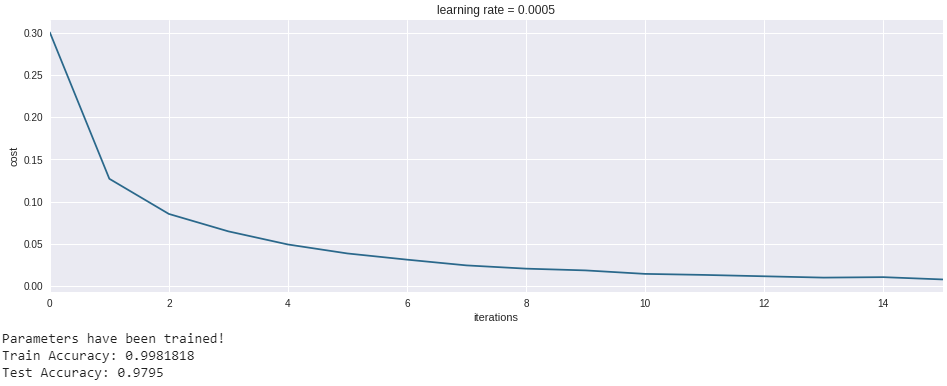


When hidden number of notes is reduced, the accuracy will be down.

(d) For the gradient optimization techniques (such as Momentum, Adagrad, ADAM, RMSProp, Nesterov Accelerated Gradient, AdaDelta, etc.), experiment on at least two diﬀerent techniques. Show the comparison results in a table. Discuss why you have the result.

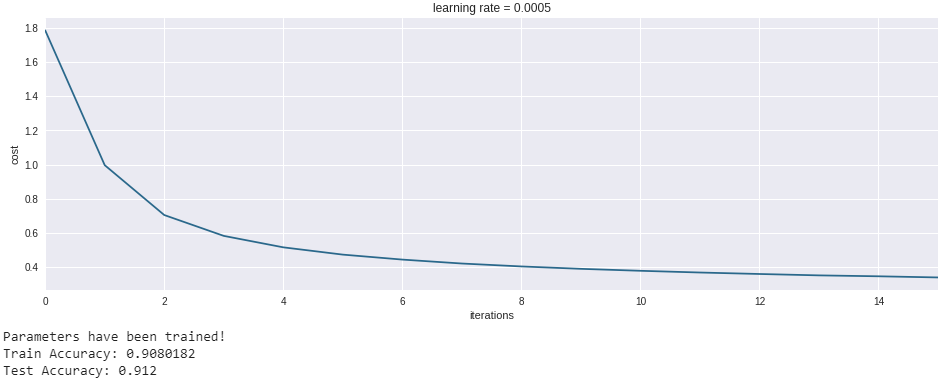
ADAM:





Adagrad:

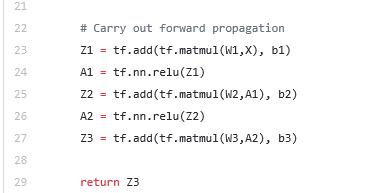
C:\Users\IAI User\AppData\Roaming\Tencent\Users\570117572\QQ\WinTemp\RichOle\SCBD((9}WI@@GF4XR9(L8(V.png

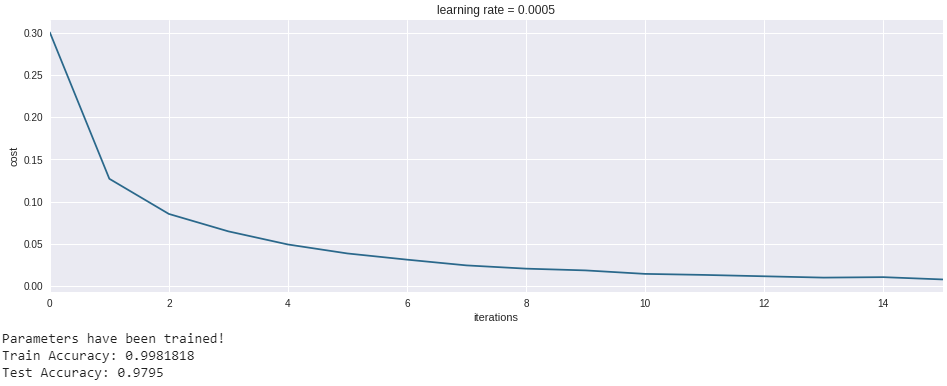


At the beginning, the Adagrad method made the cost high than ADAM, but the cost became lower with time. Adam is a different parameter adaptive learning rate method. It uses the first moment estimation and second moment estimation of the gradient to dynamically adjust the learning rate of each parameter. The main advantage of Adam is that after the offset correction, each iteration learning rate has a certain range, which makes the parameters relatively stable. For Adagrad method, when the previous period is small, the regularizer is larger and can amplify the gradient; when the later period is larger, the regularizer is smaller and can constrain the gradient; suitable for handling sparse gradients.

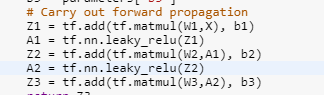
(e) For the activation functions (ReLU, Leaky ReLU, PReLU, SELU, or Swish), experiment on at least two diﬀerent functions. Show the comparison results in a table. Discuss why you have the result.

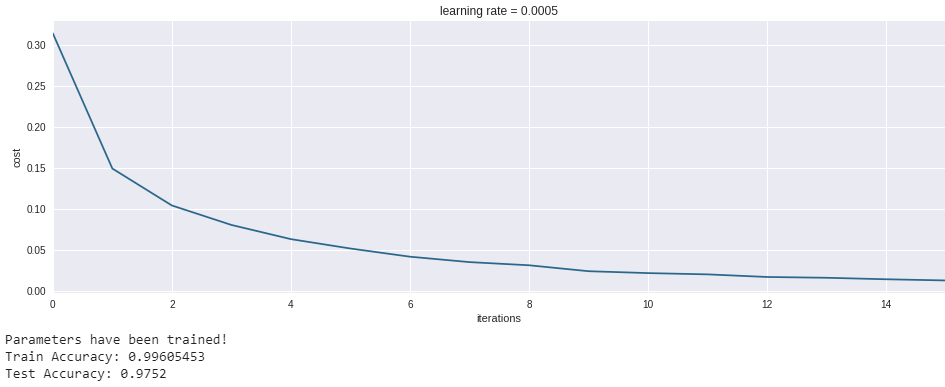
RELU：





Leaky ReLU:

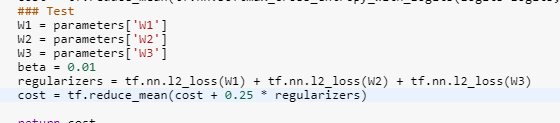


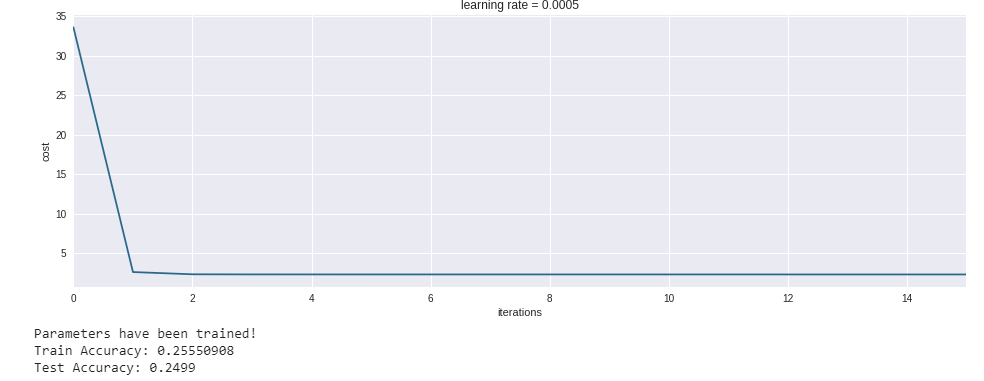


When using a good initialization of weights, ReLU, and Leaky relu perform nearly the same. Perhaps since the problem is very simple, there is no distinct advantage of learning the parameter beta.

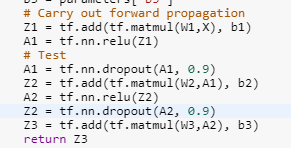
(f) For the regularization techniques (L1, L2, Dropout), experiment on at least two different techniques. Show the comparison results in a table. Discuss why you have the result.

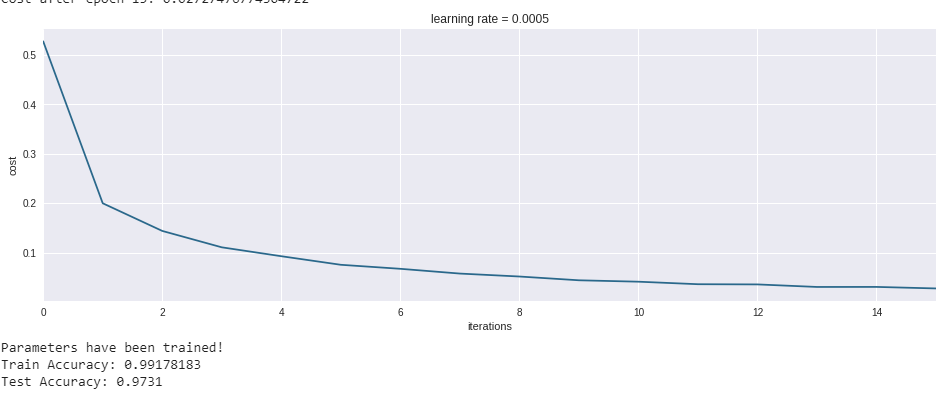
L2:





Dropout:





The result of dropout is more smoothing than L2. The L2 norm is the sum of the squares of the parameters and then the square root. Let us minimize the regular term of the L2 norm, so that each element of W is small and close to zero. But unlike the L1 norm, it does not mean that each element is 0, but only close to zero. The smaller the parameter, the simpler the model, and the simpler the model, the less likely it is to over-fitting. For dropout method, it is precisely because some units are randomly discarded at each layer, so the network equivalent to training is much smaller than the normal network, which explains to some extent the problem of avoiding over-fitting.

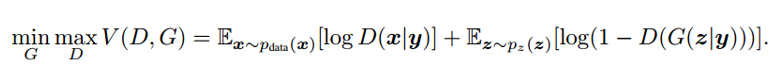
9. (This homework is due: Nov. 15) Perform GAN training on Fashion MNIST datasets 4. You can use whatever tools but TensorFlow/Python or TensorFlow/R is recommended. Or you can use MXNet R package 5. Write a detailed report (including the summary of analyzed ﬁles). One example report is as follows, but you need to write far more detailed report including the qualitative and quantitative analysis of the experimental results:

https://github.com/mari-linhares/DeepLearning/tree/master/GAN-fashion-MNIST

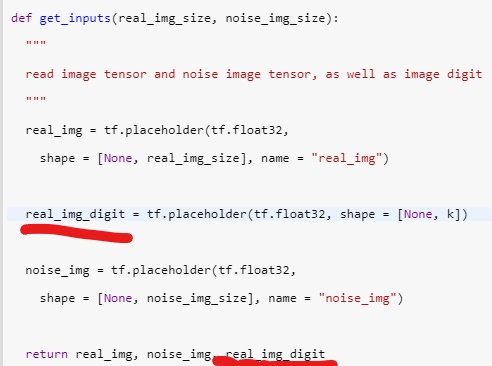
You can use any GAN algorithm 6 that might be appropriate. I personally recommended Conditional GAN. You have to save the trained model and provide a way to run the model for inference. Use a word processor (MS Word or L ATEX). L ATEXis recommended. You need to submit the report as well as screenshots. Github is recommended for project related ﬁles (PDF and PNG ﬁles) and you can simply send me an URL of Github, instead of emailing zip ﬁles.

Difference between GAN and CGAN:

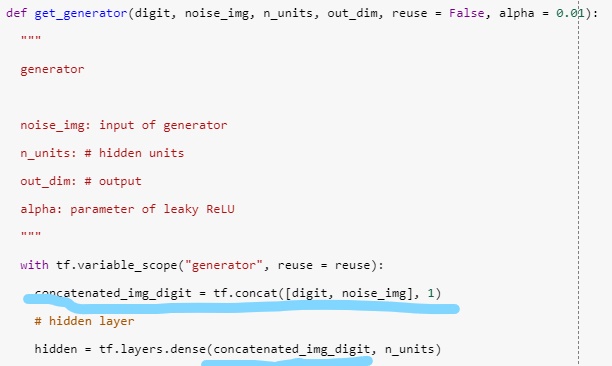
CGAN is based on GAN. We can see that from the formula below,there will be an additional label to the input of discriminator and generator.

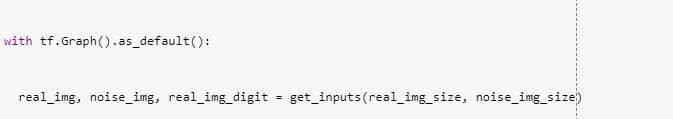


Compared with GAN, I try to find the differences in CGAN:



In the function of get\_inputs() , the label(real\_img\_digit) is also a element to input.

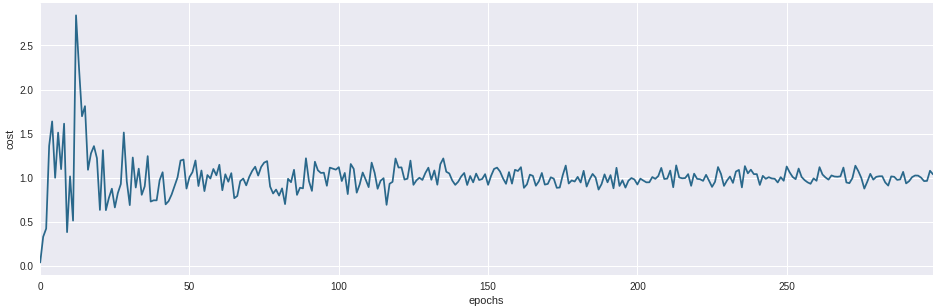




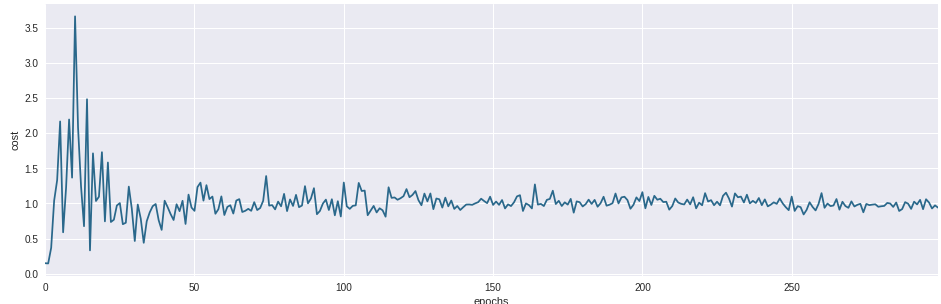
Label “digit” is added to the input of generator and discriminator.

loss:

GAN



CGAN



We can see that ,the result of CGAN is better than GAN. The reason is that, CGAN gives clearly lables to generator and discriminator. In this reason, generator will have clear definition, and have little time to generate ideal picture to cheat discriminator. For discriminator, it can discriminate fake picture from original picture efficiently. All in all, GGAN has better performance than GAN.