In The Name of God

Assignment1

Dr.Kheradpisheh

Author : Mohamadreza Khanmohamadi

Lesson : Neural Network

November 10, 2022

Problem 1

Describe in detail how L1 regularization differs from L2 regularization, and which one do you prefer?
In addition, please describe intuitively how each affects the model weights.

Solution.

Obviously one of the most important duty of the deep learning engineer is preventing model from overfitting. overfitting probablity grows as much as model complexity. if model complexity become more than data complexity we confront with overfitting. One of the approach for preventing this condition is regularization. intuitionaly, L1 try to converge the weight to zero and L2 try to reduce the weights equally. so L1 is good at feature selection and L2 is good when the features has correlation.

L1 Regularization :

$$L(x,y) \equiv \sum_{i=1}^{n} (y_i - h_\theta(x_i))^2 + \lambda \sum_{i=1}^{n} |\theta_i|$$

L1 try to reduce weight and parameters that doesnt have much importance and try to make parameters zero. at the end this soloution cause sparce zero L1.
From the above formula, we can say that:

a) When w is positive, the regularization parameter ($\lambda > 0$) will make w to be least positive, by deducting λ from w.

b) When w is negative, the regularization parameter ($\lambda < 0$) will make w to be little negative, by summing λ to w.

The regression model that uses L1 regularization technique is called Lasso Regression.

L1 Regularization :

$$L(x,y) \equiv \sum_{i=1}^{n} (y_i - h_\theta(x_i))^2 + \lambda \sum_{i=1}^{n} \theta_i^2$$

L2 force parameters to be small (Unlike L1 which force them to become zero) as much as it can. L2 suffers the outliers because the outliers predict has high error and loss then force the weights to be smaller. so L2 works well when the outliers disappear and data has correlation features.

Lets Compare (L1 Vs L2):

a) Panelizes the sum of absolute value of weights vs penalizes the sum of square weights.

b) It has a sparse solution. It has a non-sparse solution.

c) It gives multiple solutions. It has only one solution

d) Constructed in feature selection. No feature selection

e) Constructed in feature selection. No feature selection

f) Robust to outliers. Not robust to outliers.

g) It generates simple and interpretable models.It gives more accurate predictions when the output variable is the function of whole input variables.

h) Unable to learn complex data patterns. Able to learn complex data patterns.

Problem 2

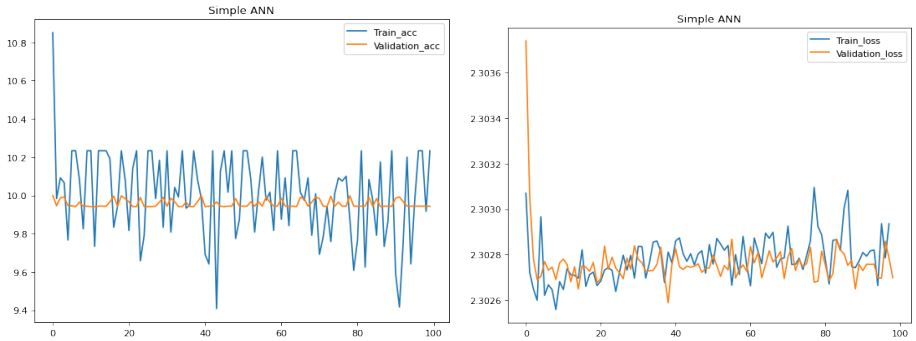
In this part, you will implement a simple multi-layer perceptron neural network using PyTorch to solve a clothing classification problem. You have to work with the Fashion MNIST dataset, which consists of 10 classes with 60,000 examples in the training set and 10,000 examples in the test set.

Solution.

1 Models

1.1 First Model

At first model we use this architecture :
RELU(Linear(784,512))
RELU(Linear(512,64))
RELU(Linear(64,10))
epoch : 100
Learning rate : 1e -2
optimizer : Adam

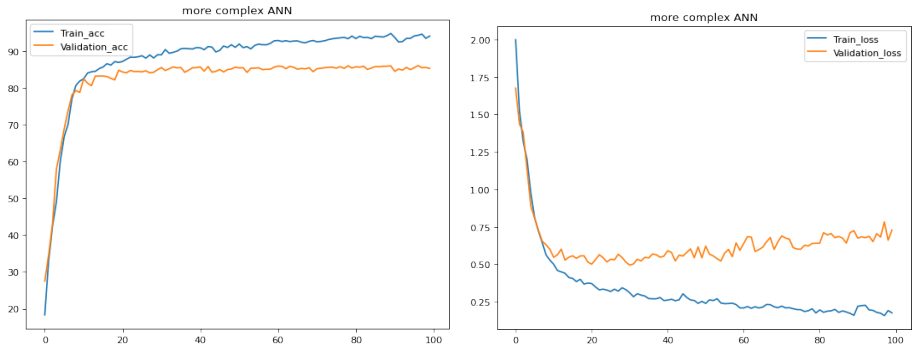


The loss and accuracy on the train and set was very bad. loss didnt deacrease and val loss doesnt have any change. this is in confilct of overfitting. this model is so simple so it is underfit.

1.2 Second Model

we use more complex model :
RELU(Linear(784,512))
RELU(Linear(512,256))
RELU(Linear(256,128))
RELU(Linear(128,64))
RELU(Linear(64,32))
RELU(Linear(32,16))
RELU(Linear(16,10))

epoch : 100
Learning rate : 1e-2
optimizer : Adam



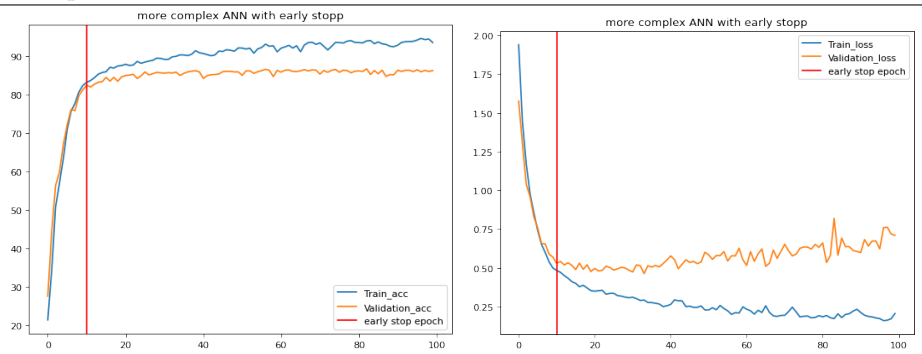
The loss and accuraccy is good in this model. but after some epochs there is no significant change in the loss and accuracy.

1.3 Early stopping Model

Early stopping is a form of regularization used to avoid overfitting when training a learner with an iterative method, such as gradient descent. we use this method and implement it as :

when the epsilon $\leq$ (validation loss - train loss)

then we count these situations. if the counter more than telorance then the iteration and epoch will break. lets see the results :

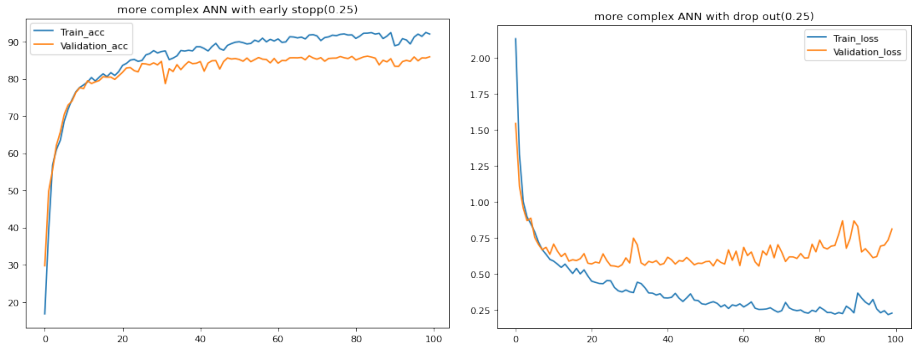


as we see on the epoch 10 the training stopped. it is obvious that in 10th iteration the loss and acc in valid and train become closer and it is optimum epoch to stop.

1.4 Dropout Model

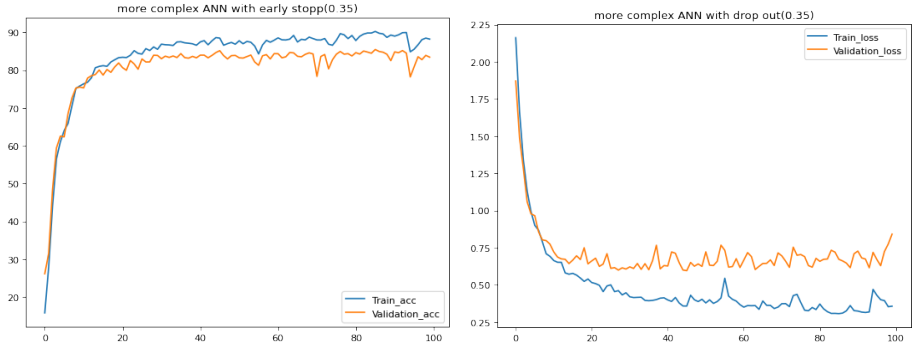
Lets see the architecture :
RELU(Linear(784,512))
RELU(Linear(512,256))
Dropout(0.25)
RELU(Linear(256,128))
RELU(Linear(128,64))
Dropout(0.25)
RELU(Linear(64,32))
RELU(Linear(32,16))
Dropout(0.25)
RELU(Linear(16,10))

epoch : 100
$\text{Learning}_r ate : 1e-2$
$optimizer : Adam$



Dropout is a machine learning technique where you remove (or "drop out") units in a neural net to simulate training large numbers of architectures simultaneously. Importantly, dropout can drastically reduce the chance of overfitting during training. As we see, we can see that the model with drop out trick overfitted in later epochs . if the dropout parameter

change to 0.5 it become more clear.



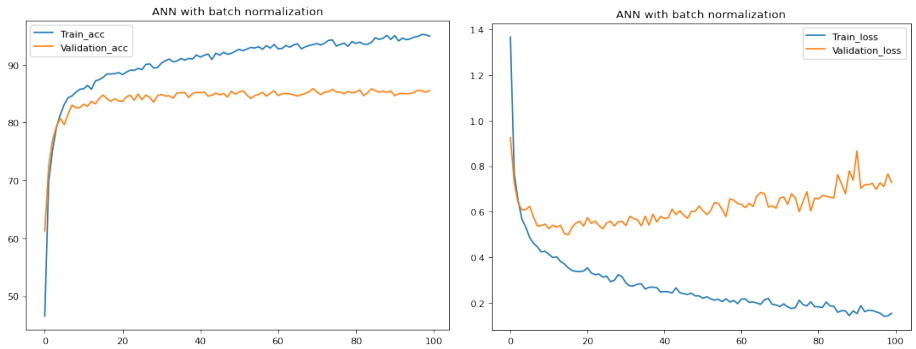
1.5 BatchNormalization Model

Batch normalization acts to standardize only the mean and variance of each unit in order to stabilize learning, but allows the relationships between units and the nonlinear statistics of a single unit to change.By only using Batch Normalization, we match the accuracy of Inception in less than half the number of training steps.

Lets see the architecture :

RELU(Linear(784,512))
batchNorm1d(RELU(Linear(512,256)))
RELU(Linear(128,64))
batchNorm1D(RELU(Linear(64,32)))
RELU(Linear(32,16))
RELU(Linear(16,10))

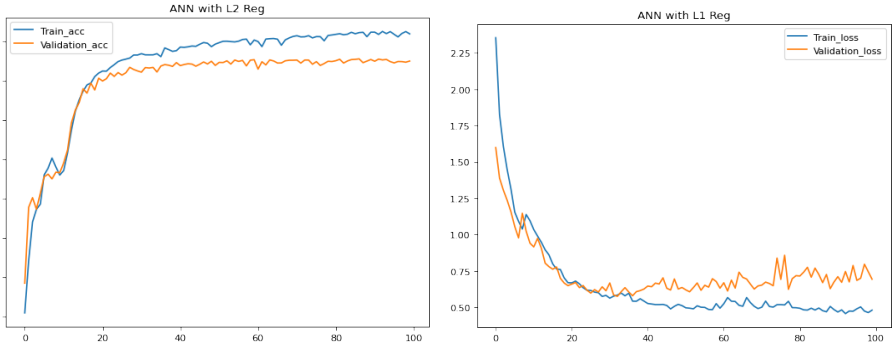
epoch : 100
Learning rate : 1e-2
optimizer : Adam



The results show that the convergence of the model become faster and model overfit in lower epochs. This mean that if we use the trick of overfitting the model get better accuracy and lower loss.

1.6 L1 Regularization

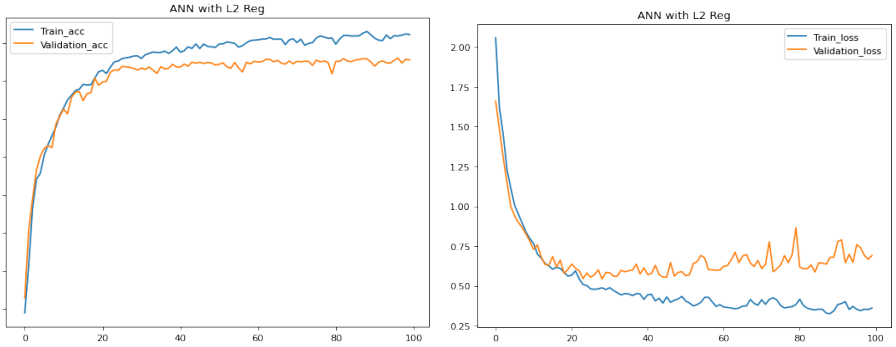
As i mentioned later, the L1 try to remove unimportant nodes and select best and important nodes which can detect important features. lets see results :



The results approve my speechs. the model got lower loss and get better accuracy in comparison with basic model.

1.7 L2 Regularization

Lets see the L2 effect on the basic model :



As we the see results (loss and accuracy), it shows that the accuracy and loss in the valid data is better in compare with the basic model.