Introduction to Deep Learning CS6910

Assignment 2

Submitted by
Chella Thiyagarajan N
ME17B179

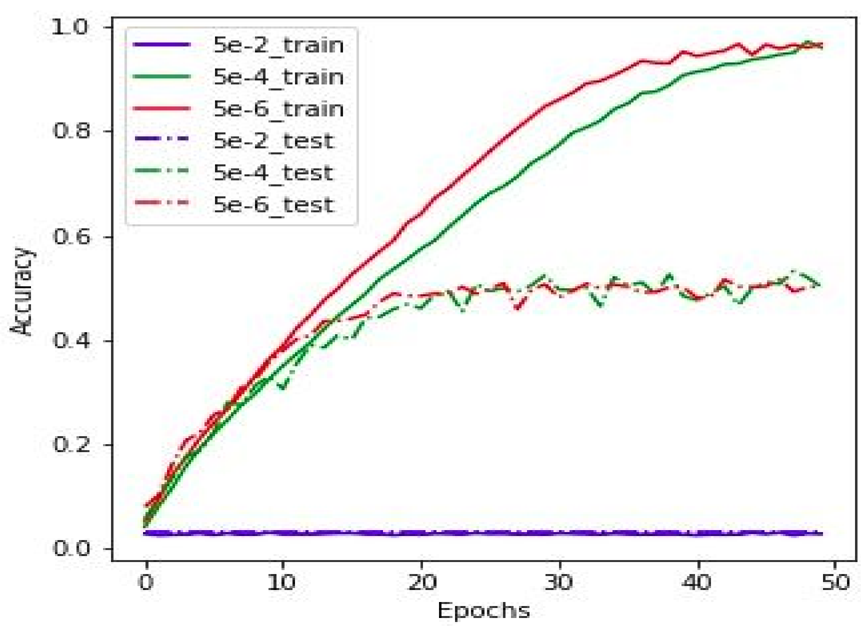
Part A:

Aim: Take the best performing model from Assignment-1. Experiment with various regularization parameter (by changing the weight decay parameter of the optimizer) values and find the best one. Visualize the train vs test accuracy, loss for each value.

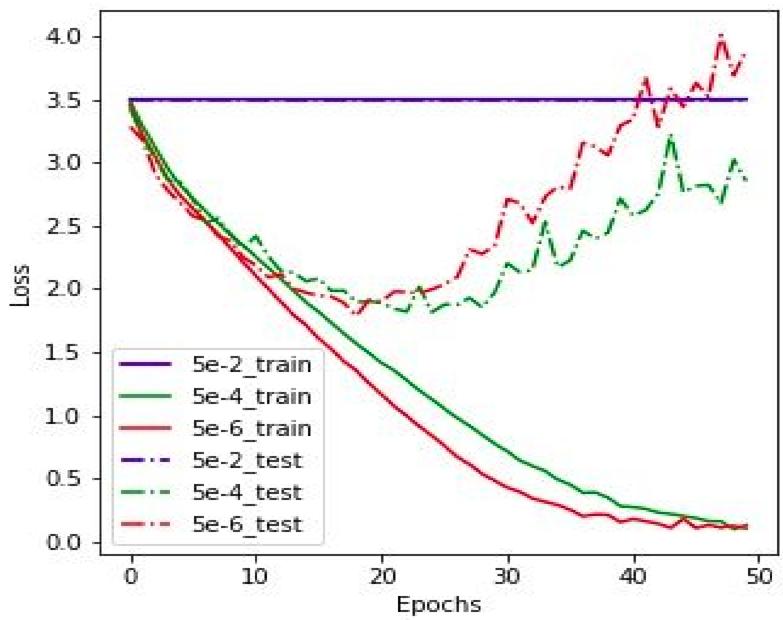
Overview:
This section looks at the observations when the weight decay regularisation parameters are changed from the Best model obtained from assignment 1. Note that the other features of the model across all the layers are kept constant during the study. It has been experimented with 5e-2, 5e-4, and 5e-6 as it's weight decay parameter in stochastic gradient descent function.

Results:

Train and Test Accuracy vs Epochs Plot



Train and Test Loss vs Epochs Plot



Inferences:

From the above plots, we can infer that a model with weight decay of 5e-2 does not learn anything as we progress through the epochs. 5e-2 is too large a value for the model to converge for our particular dataset. It is reflected in the above plots as the blue line represented by 5e-2 is flat with no changes.

Our model with weight decay parameter 5e-6 tends to dominate the 5e-4 model in terms of training accuracies across epochs, but once the number of epochs is high enough they converge to the same value. The same trend can be observed with train losses too, as the 5e-6 model has low loss compared to the 5e-4 model but after a certain number of epochs, they tend to converge.

Both 5e-6 and 5e-4 models show the same test accuracies throughout all the epochs. When we observe the test losses of model 5e-4 and 5e-6, we can observe after epoch number 23 their losses start to increase when it is actually expected to decrease. This may be due to overfitting. In the increasing phase after 23 epochs, 5e-6 model losses increase rapidly than 5e-6 losses.

Therefore we can conclude that with a given reasonably high number of epochs 5e-4 model performs better than other models.

Part - B

D. Gradient Calculation of Common Activation functions

a) Sigmoid

$$f(x) = \frac{1}{1+e^{-x}}$$

$$\frac{df}{dx} = \frac{d}{dx}\left[(1+e^{-x})^{-1}\right]$$

Let $1+e^{-x} = g(x)$

$$\frac{df}{dx} = \frac{d}{dx}\left[g(x)^{-1}\right] = \frac{-1}{(g(x))^2}\frac{dg(x)}{dx}$$

$$= \frac{-1}{(1+e^{-x})^2}\left[-e^{-x}\right] = \frac{e^{-x}}{(1+e^{-x})^2}$$

Decomposition :

$$\frac{df}{dx} = \frac{1}{1+e^{-x}} - \frac{1}{(1+e^{-x})^2} = \frac{1}{1+e^{-x}}\left[1 - \frac{1}{1+e^{-x}}\right]$$

$$\therefore \boxed{\frac{df}{dx} = (f(x))(1-f(x))} \longrightarrow 4$$

b) Hyperbolic Tangent

$$f(x) = \tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

Let $g_1(x) = e^{x} - e^{-x}$

$g_2(x) = e^{x} + e^{-x}$

$$\frac{df}{dx} = \frac{d}{dx}\left[\frac{g_1(x)}{g_2(x)}\right]$$

$$= \frac{1}{g_2(x)} \frac{d}{dx}\left[g_1(x)\right] - \frac{g_1(x)}{(g_2(x))^2} \frac{d}{dx}\left[g_2(x)\right] \longrightarrow (1)$$

$$\frac{d}{dx} g_2(x) = e^x - e^{-x} = g_1(x) \longrightarrow (2)$$

$$\frac{d}{dx} g_1(x) = e^x + e^{-x} = g_2(x) \longrightarrow (3)$$

Decomposition :

use (2) & (3) in (1)

$$\frac{df}{dx} = \frac{g_2}{g_2} - \frac{g_1}{g_2^2} \cdot g_1 = 1 - \frac{g_1^2}{g_2^2}$$

$$= 1 - \left[\frac{g_1}{g_2}\right]^2 = 1 - (f(x))^2$$

$$\boxed{\frac{df}{dx} = 1 - (f(x))^2} \longrightarrow (5)$$

(iii) ReLu

$$f(x) = \begin{cases} x, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

$$\boxed{\frac{df}{dx} = \begin{cases} 1, & x > 0 \\ 0, & x \leq 0 \end{cases}} \longrightarrow (7)$$

There is no decomposition or simplification step here. $\frac{df}{dx}$ is a piece-wise operation

2. Gradient Calculation of Common loss functions

a) Cross Entropy Loss

$$f(x, \theta, y) = -\sum_i y_i(x_i) P_\theta(x_i)$$

here, θ is model parameters

$P_\theta(x_i)$ is predicted value

$y_i(x_i)$ is true value

x is input

$$\frac{\partial f}{\partial \theta} = -\sum_i y_i(x_i) \frac{\partial (P_\theta(x_i))}{\partial \theta}$$

Case 1: Let $P_\theta(x_i)$ be a sigmoid function

$$\frac{\partial f}{\partial \theta} = -\sum_i y_i(x_i) \frac{\partial}{\partial \theta}\left[\frac{1}{1+e^{-x_i\theta}}\right] \longrightarrow (6)$$

from (4) & (6)

$$\frac{\partial f}{\partial \theta} = -\sum_i y_i(x_i)\left[\frac{1}{1+e^{-x_i\theta}}\right]\left[1 - \frac{1}{1+e^{-x_i\theta}}\right]\frac{\partial}{\partial \theta}[x_i \cdot \theta]$$

$$\boxed{\frac{\partial f}{\partial \theta} = -\sum y_i(x_i)\left[\frac{1}{1+e^{-x_i\theta}}\right]\left[1 - \frac{1}{1+e^{-x_i\theta}}\right] x_i}$$

Case 2: Let $P_\theta(x_i)$ be a hyperbolic Tangent

from (5) & (6)

$$\boxed{\frac{\partial f}{\partial \theta} = -\sum_i y_i(x_i)\left(1 - \tanh^2(x_i\theta)\right) x_i}$$

Case 3: Let $P_\theta(x_i)$ be a ReLu function

from (7) & (6)

$$\boxed{\frac{\partial f}{\partial \theta} = \begin{cases} -\sum_i y_i, & x > 0 \\ 0, & x \leq 0 \end{cases}}$$

b) Hinge Loss [y_i = Correct class of sample :]

$$f(x,\theta,y) = \max(0, \cancel{1}\, y \cdot f_\theta(x))$$

$$f = \begin{cases} 1 - y \cdot f_\theta(x), & y f_\theta(x) < 1 \\ 0, & y f_\theta(x) \geq 1 \end{cases}$$

$$\frac{\partial f}{\partial \theta} = \begin{cases} -y \frac{\partial}{\partial \theta}(f_\theta(x)), & y f_\theta(x) < 1 \\ 0, & y f_\theta(x) \geq 1 \end{cases}$$

Let $g(x,\theta,y) = -y \frac{\partial}{\partial \theta}(f_\theta(x))$

Case 1 : Sigmoid function

$$g(x,\theta,y) = \sum_i -y_i \left[\frac{1}{1-e^{-x_i\theta}}\right]\left[1 - \frac{1}{1-e^{-x_i\theta}}\right] x_i$$

Case 2 : Hyperbolic Tangent

$$g = -\sum_i y_i \left(1 - \tanh^2(x_i\theta)\right) x_i$$

Case 3 : Relu

$$g = \begin{cases} -\sum_i y_i, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

8

c) L_1 loss

$$f(x,\theta,y) = \sum_i |y_i(x_i) - f_\theta(x_i)|$$

$$\frac{\partial f}{\partial \theta} = \frac{\partial}{\partial \theta}\left[\sum_i |y_i(x_i) - f_\theta(x_i)|\right]$$

$$\frac{\partial f}{\partial \theta} = \begin{cases} -\sum_i \frac{\partial}{\partial \theta} f_\theta(x_i), & y_i(x_i) > f_\theta(x_i) \\ \sum_i \frac{\partial}{\partial \theta} f_\theta(x_i), & y_i(x_i) < f_\theta(x_i) \end{cases}$$

In the above equation let $g(x,\theta,y) = \frac{\partial}{\partial \theta} f_\theta(x_i)$ which is similar to equation 8 depending on it's activation function

d) Huber Loss

$$f(x,\theta,y) = \sum_i \begin{cases} \frac{1}{2}(y_i(x_i) - f_\theta(x_i))^2, & |y_i - f_\theta(x_i)| \leq \delta \\ \delta\, |y_i - f_\theta(x_i)| - \frac{\delta^2}{2}, & |y_i - f_\theta(x_i)| > \delta \end{cases}$$

$$\frac{\partial f}{\partial \theta} = \sum_i \begin{cases} (y_i - f_\theta(x_i)) \frac{\partial}{\partial \theta}[f_\theta(x_i)], & |y_i - f_\theta(x_i)| \leq \delta \\ -\delta \frac{\partial}{\partial \theta}[f_\theta(x_i)], & |y_i - f_\theta(x_i)| > \delta \text{ and } y_i > f_\theta(x_i) \\ \delta \frac{\partial}{\partial \theta}[f_\theta(x_i)], & |y_i - f_\theta(x_i)| > \delta \text{ and } y_i < f_\theta(x_i) \end{cases}$$

In the above equation let $g(x,\theta,y) = \frac{\partial}{\partial \theta} f_\theta(x_i)$ which is similar to equation 8 depending on it's activation function

e) L_2 loss

$$f(x, \theta, y) = \sum_i \left(y_i(x_i) - f_\theta(x_i)\right)^2$$

$$\frac{\partial f}{\partial \theta} = 2 \sum_i \left(y_i - f_\theta(x_i)\right) \cdot \frac{\partial f_\theta(x_i)}{\partial \theta}$$

In the above equation Let $g(x, \theta, y) = \frac{\partial}{\partial \theta}\left[f_\theta(x_i)\right]$ which is similar to equation (8) dependending on it's activation function

f) Cosine Similarity

$$f(x, \theta, y) = \sum_i \left[1 - \frac{y_i^T f_\theta(x_i)}{\|y_i\| \cdot \|f_\theta(x_i)\|} \right]$$

$$\frac{\partial f}{\partial \theta} = \sum_i \left[\frac{-y_i^T \cdot \frac{\partial}{\partial \theta} f_\theta(x_i)}{\|y_i\| \left[f_\theta^T \cdot f_\theta\right]^{1/2}} + \frac{y_i^T f_\theta(x_i)}{2 \|y_i\| \left(f_\theta^T \cdot f_\theta\right)^{3/2}} \left[\frac{\partial}{\partial \theta} f^T f + f_\theta^T \ldots \right. \right.$$

In the above equation Let $g(x, \theta, y) = \frac{\partial}{\partial \theta}\left[f_\theta(x_i)\right]$ which is similar to equation (8) depending on it's activation function.

3) Hand - Calculation of gradients

inp 1 = -1.67   target = -0.49 ω_{12-1} = -0.15

inp 2 = 0.98 ω_{11-1} = 0.39 ω_{22-1} = 0.49

inp 3 = -0.71 ω_{21-1} = 0.47 ω_{32-1} = 0.19

ω_{31-1} = 0.06

$B1_{-1}$ = -0.01 ω_{11-2} = -0.01 B_{1-2} = 0.11

$B2_{-1}$ = -0.57 ω_{21-2} = 0.33

$$f_\theta(x) = \omega_{11-2}\,\sigma(\omega_{11-1} i_1 + \omega_{21-1} i_2 + \omega_{31-1} i_3) + \omega_{21-2}\,\sigma[\omega_{12-1} i_1 + \omega_{22-1} i_2 + \omega_{32-1} i_3]$$

here i_1, i_2 & i_3 are inputs

$$L_\theta(x,y) = [\text{target} - f_\theta(x)]^2$$

$$\frac{\partial L_\theta}{\partial \omega_{abc}} = 2[\text{target} - f_\theta(x)] \cdot -\frac{\partial f_\theta(x)}{\partial \omega_{abc}}$$

Weight update step:

$$\omega_{ab-c} = \omega_{ab-c} - \frac{l \cdot \partial L_\theta(x,y)}{\partial \omega_{ab-c}}$$

here l is learning rate

$$\frac{\partial L}{\partial \omega_{11-2}} = B_{1-1} \times -2(\text{target} - B_{1-2}) = -0.01 \times 1.2 = -0.012$$

$$\frac{\partial L}{\partial \omega_{21-2}} = B_{2-1} \times -2(\text{target} - B_{1-2}) = -0.57 \times 1.2 = -1.77$$

$$\frac{\partial L}{\partial \omega_{11-1}} = \omega_{11-2} \cdot B_{1-1} \cdot (1 - B_{1-1}) i_1 \times -2(\text{target} - B_{1-2}) = 0.00657 \times 1.2 = 0.007884$$

$$\frac{\partial L}{\partial \omega_{21\text{-}1}} = \omega_{11\text{-}2} \cdot B_{1\text{-}1} \cdot (1 - B_{1\text{-}1})\, i_2 \times -2(\text{target} - B_{1\text{-}2}) = 0.000098 \times 1.2 = 0.0001176$$

$$\frac{\partial L}{\partial \omega_{31\text{-}1}} = \omega_{11\text{-}2} \cdot B_{1\text{-}1} \cdot (1 - B_{1\text{-}1})\, i_3 \times -2(\text{target} - B_{1\text{-}2}) = -0.00007171 \times 1.2 = -0.000086$$

$$\frac{\partial L}{\partial \omega_{12\text{-}1}} = \omega_{21\text{-}2} \cdot B_{2\text{-}1} \cdot (1 - B_{2\text{-}1})\, i_1 \times -2(\text{target} - B_{1\text{-}2}) = -0.2953 \times -1.67 = 0.493 \times 1.2 = 0.5916$$

$$\frac{\partial L}{\partial \omega_{22\text{-}1}} = \omega_{21\text{-}2} \cdot B_{2\text{-}1} \cdot (1 - B_{2\text{-}1})\, i_2 \times -2(\text{target} - B_{1\text{-}2}) = -0.2893 \times 1.2 = -0.34716$$

$$\frac{\partial L}{\partial \omega_{32\text{-}1}} = \omega_{21\text{-}2} \cdot B_{2\text{-}1} \cdot (1 - B_{2\text{-}1})\, i_3 \times -2(\text{target} - B_{1\text{-}2}) = 0.2096 \times 1.2 = 0.25152$$

Let learning rate (ℓ) = 0.1

~~0.01~~

updated weights :

$\omega_{11\text{-}2} = \cancel{-0.009}\quad -0.0088$

$\omega_{21\text{-}2} = \cancel{0.387}\quad 0.507$

$\omega_{11\text{-}1} = \cancel{0.389343}\quad 0.392116$

$\omega_{21\text{-}1} = \cancel{0.46999}\quad 0.469988$

$\omega_{31\text{-}1} = \cancel{0.06007171}\quad 0.06000859$

$\omega_{12\text{-}1} = \cancel{-0.1993}\quad -0.20916$

$\omega_{22\text{-}1} = \cancel{0.51893}\quad 0.524716$

$\omega_{32\text{-}1} = \cancel{0.16904}\quad 0.164848$