Deep Learning - HW1

Afek Adler & Or Wolkomir

November 25, 2019

1 Theory

1.1 Question 1 - network architecture

(a) The input shape is (m,10)

(b) W_h shape is $(10, 50)$ and b_h shape is $(, 50)$

(c) W_o shape is $(50, 3)$ and b_0 shape is $(, 3)$

(d) y shape is $(m, 3)$

(e) W_h shape is $(50, 3)$ and b_h shape is $(, 3)$

1.2 Question 2 - # of parameters

(a) **Conv layer 1**
Kernel = 3
Channels = 3
Output = 100
Bias := Output
Therefore:

$$\#\_of\_parameters = Kernel^2 * Channels * Output + Bias \Rightarrow$$
$$3^3 * 100 + 100 = 2,800$$

(b) **Conv layer 2**
Kernel = 3
Channels = 100
Output = 200
Bias := Output
Therefore:

$$\#\_of\_parameters = Kernel^2 * Channels * Output + Bias \Rightarrow$$
$$3^2 * 100 * 200 + 200 = 180,200$$

(c) **Conv layer3**
Kernel = 3
Channels = 200
Output = 400
Bias := Output
Therefore:

$$\#\_of\_parameters = Kernel^2 * Channels * Output + Bias \Rightarrow$$
$$3^2 * 200 * 400 + 400 = 720,400$$

(d) **Total**

$$total\_\#\_of\_parameters = 720,400 + 180,200 + 2,800 = 903,400$$

1.3 Question 3 - deriving BatchNorm gradients

(a) $\frac{\partial f}{\partial \gamma} =$

$$\frac{\partial f}{\partial y} \cdot \frac{\partial y}{\partial \gamma} = \sum_{i=1}^{m} \frac{\partial f}{\partial y_i} \cdot \widehat{x}_i \tag{1}$$

(b) $\frac{\partial f}{\partial \beta} =$

$$\frac{\partial f}{\partial y} \cdot \frac{\partial y}{\partial \beta} = \sum_{i=1}^{m} \frac{\partial f}{\partial y_i} \frac{\partial f}{\partial \beta} = \sum_{i=1}^{m} \frac{\partial f}{\partial y_i} \tag{2}$$

(c) $\frac{\partial f}{\partial \widehat{x}_i} =$

$$\frac{\partial f}{\partial y} \cdot \frac{\partial y}{\partial \widehat{x}_i} = \frac{\partial f}{\partial y_i} \cdot \gamma \tag{3}$$

(d) $\frac{\partial f}{\partial \sigma^2} =$

$$\frac{\partial f}{\partial \widehat{x}} \cdot \frac{\partial \widehat{x}}{\partial \sigma^2} = \frac{\partial y}{\partial \widehat{x}_i} = \frac{\partial f}{\partial y_i} \cdot \gamma \tag{4}$$

(e) $\frac{\partial f}{\partial \mu} =$

$$\frac{\partial f}{\partial \mu} = \frac{\partial f}{\partial \hat{x}} \cdot \frac{\partial \hat{x}}{\partial \mu} + \frac{\partial f}{\partial \sigma^2} \cdot \frac{\partial \sigma^2}{\partial \mu} =$$
$$= \sum_{i=1}^{m} \frac{\partial f}{\partial \hat{x}_i} \cdot \frac{\partial \hat{x}_i}{\partial \mu} + \frac{\partial f}{\partial \sigma^2} \cdot \frac{\partial \sigma^2}{\partial \mu}$$
$$\frac{\partial \hat{x}_i}{\partial \mu} = \frac{-1}{\sqrt{\sigma^2+\epsilon}}$$
$$\frac{\partial \sigma^2}{\partial \mu} = -\frac{2}{m} \cdot \sum_{i=1}^{m} (x_i - \mu)$$
$$\frac{\partial f}{\partial \sigma^2} = \left(-\frac{1}{2}\right) \cdot \sum_{i=1}^{m} \frac{\partial f}{\partial \hat{x}_i} \cdot (x_i - \mu) \cdot \left(\sigma^2 + \epsilon\right)^{-\frac{3}{2}} \frac{\partial f}{\partial \mu} = \left(\sum_{i=1}^{m} \frac{\partial f}{\partial \hat{x}_i} \cdot \frac{\partial \hat{x}_i}{\partial \mu}\right) + \left(\frac{\partial f}{\partial \sigma^2} \cdot \frac{\partial \sigma^2}{\partial \mu}\right) =$$
$$= \left(\sum_{i=1}^{m} \frac{\partial f}{\partial \hat{x}_i} \cdot \frac{-1}{\sqrt{\sigma^2+\epsilon}}\right) + \left(\frac{\partial f}{\partial \sigma^2} \cdot \left(-\frac{2}{m}\right) \cdot \sum_{i=1}^{m} (x_i - \mu)\right) =$$
$$= \sum_{i=1}^{m} \frac{\partial f}{\partial \hat{x}_i} \cdot \frac{-1}{\sqrt{\sigma^2+\epsilon}} + \frac{\partial f}{\partial \sigma^2} \cdot \left(-\frac{2}{m}\right) \cdot [\sum_{i=1}^{m} (x_i) - \sum_{i=1}^{m} (\mu)] = \sum_{i=1}^{m} \frac{\partial f}{\partial \hat{x}_i} \cdot \frac{-1}{\sqrt{\sigma^2+\epsilon}}$$
$$+ \frac{\partial f}{\partial \sigma^2} \cdot (-2) \cdot \left[\frac{1}{m} \sum_{i=1}^{m} (x_i) - \frac{1}{m} \sum_{i=1}^{m} (\mu)\right] = = \sum_{i=1}^{m} \frac{\partial f}{\partial \hat{x}_i} \cdot \frac{-1}{\sqrt{\sigma^2+\epsilon}} + \frac{\partial f}{\partial \sigma^2} \cdot (-2) \cdot \left[\mu - \frac{1}{m} \sum_{i=1}^{m} \mu\right] =$$
$$= \sum_{i=1}^{m} \frac{\partial f}{\partial \hat{x}_i} \cdot \frac{-1}{\sqrt{\sigma^2+\epsilon}} + \frac{\partial f}{\partial \sigma^2} \cdot (-2) \cdot \left[\mu - \frac{1}{m} \cdot m \cdot \mu\right] =$$
$$= \sum_{i=1}^{m} \frac{\partial f}{\partial \hat{x}_i} \cdot \frac{-1}{\sqrt{\sigma^2+\epsilon}} + \frac{\partial f}{\partial \sigma^2} \cdot 0 =$$
$$= \sum_{i=1}^{m} \frac{\partial f}{\partial \hat{x}_i} \cdot \frac{-1}{\sqrt{\sigma^2+\epsilon}} \tag{5}$$

(f) $\frac{\partial f}{\partial x_i} =$

$$\frac{\partial f}{\partial x_i} = \frac{\partial f}{\partial \hat{x}_i} \cdot \frac{\partial \hat{x}_i}{\partial x_i} + \frac{\partial f}{\partial \mu} \cdot \frac{\partial \mu}{\partial x_i} + \frac{\partial f}{\partial \sigma^2} \cdot \frac{\partial \sigma^2}{\partial x_i} =$$
$$\frac{\partial \hat{x}_i}{\partial x_i} = \frac{1}{\sqrt{\sigma^2+\epsilon}}$$
$$\frac{\partial \mu}{\partial x_i} = \frac{1}{m}$$
$$\frac{\partial \sigma^2}{\partial x_i} = \frac{2}{m} \cdot (x_i - \mu) \frac{\partial f}{\partial x_i} = \frac{\partial f}{\partial \hat{x}_i} \cdot \frac{\partial \hat{x}_i}{\partial x_i} + \frac{\partial f}{\partial \mu} \cdot \frac{\partial \mu}{\partial x_i} + \frac{\partial f}{\partial \sigma^2} \cdot \frac{\partial \sigma^2}{\partial x_i} =$$
$$= \left(\frac{\partial f}{\partial \hat{x}_i} \cdot \frac{1}{\sqrt{\sigma^2+\epsilon}}\right) + \left(\frac{\partial f}{\partial \mu} \cdot \frac{1}{m}\right) + \left(\frac{\partial f}{\partial \sigma^2} \cdot \frac{2}{m} (x_i - \mu)\right) =$$
$$= \left(\frac{\partial f}{\partial \hat{x}_i} \cdot \frac{1}{\sqrt{\sigma^2+\epsilon}}\right) + \left(\frac{1}{m} \cdot \sum_{i=1}^{m} \frac{\partial f}{\partial \hat{x}_i} \cdot \frac{-1}{\sqrt{\sigma^2+\epsilon}}\right) + \left(\frac{\partial f}{\partial \sigma^2} \cdot \frac{2}{m} (x_i - \mu)\right) =$$
$$\left(\frac{\partial f}{\partial \hat{x}_i} \cdot \frac{1}{\sqrt{\sigma^2+\epsilon}}\right) + \left(\frac{1}{m} \cdot \sum_{i=1}^{m} \frac{\partial f}{\partial \hat{x}_i} \cdot \frac{-1}{\sqrt{\sigma^2+\epsilon}}\right) + \left(\left[\left(-\frac{1}{2}\right) \cdot \sum_{j=1}^{m} \frac{\partial f}{\partial \hat{x}_j} \cdot (x_j - \mu) \cdot \left(\sigma^2 + \epsilon\right)^{-\frac{3}{2}}\right] \cdot \frac{2}{m} (x_i - \mu)\right)$$
$$= \left(\frac{\partial f}{\partial \hat{x}_i} \cdot \frac{1}{\sqrt{\sigma^2+\epsilon}}\right) + \left(\frac{1}{m} \cdot \sum_{i=1}^{m} \frac{\partial f}{\partial \hat{x}_i} \cdot \frac{-1}{\sqrt{\sigma^2+\epsilon}}\right) + \left(-\left[\frac{1}{m} \cdot \sum_{j=1}^{m} \frac{(x_j - \mu)}{\sqrt{\sigma^2+\epsilon}} \cdot \left(\sigma^2 + \epsilon\right)^{-1} \frac{\partial f}{\partial \hat{x}_j}\right] \cdot (x_i - \mu)\right)$$
$$= \left(\frac{\partial f}{\partial \hat{x}_i} \cdot \frac{1}{\sqrt{\sigma^2+\epsilon}}\right) + \left(\frac{1}{m} \cdot \sum_{i=1}^{m} \frac{\partial f}{\partial \hat{x}_i} \cdot \frac{-1}{\sqrt{\sigma^2+\epsilon}}\right) + \left(-\left[\frac{1}{m} \cdot \sum_{j=1}^{m} \frac{(x_j - \mu)}{\sqrt{\sigma^2+\epsilon}} \frac{\partial f}{\partial \hat{x}_j} \cdot \left(\sigma^2 + \epsilon\right)^{-\frac{1}{2}}\right] \cdot \frac{(x_i - \mu)}{\sqrt{\sigma^2+\epsilon}}\right)$$
$$= \left(\frac{\partial f}{\partial \hat{x}_i} \cdot \frac{1}{\sqrt{\sigma^2+\epsilon}}\right) + \left(\frac{1}{m} \cdot \sum_{i=1}^{m} \frac{\partial f}{\partial \hat{x}_i} \cdot \frac{-1}{\sqrt{\sigma^2+\epsilon}}\right) + \left(-\left[\frac{1}{m} \cdot \sum_{j=1}^{m} \hat{x}_j \cdot \frac{\partial f}{\partial \hat{x}_j} \cdot \left(\sigma^2 + \epsilon\right)^{-\frac{1}{2}}\right] \cdot \hat{x}_i\right)$$
$$= \left(\frac{\partial f}{\partial \hat{x}_i} \cdot \frac{1}{\sqrt{\sigma^2+\epsilon}}\right) + \left(\frac{1}{m} \cdot \sum_{i=1}^{m} \frac{\partial f}{\partial \hat{x}_i} \cdot \frac{-1}{\sqrt{\sigma^2+\epsilon}}\right) + \left(-\frac{1}{m \cdot \sqrt{\sigma^2+\epsilon}} \cdot \sum_{j=1}^{m} \left(\hat{x}_j \cdot \frac{\partial f}{\partial \hat{x}_j}\right) \cdot \hat{x}_i\right)$$
$$= \frac{1}{\sqrt{\sigma^2+\epsilon}} \cdot \left[\frac{\partial f}{\partial \hat{x}_i} - \frac{1}{m} \sum_{i=1}^{m} \frac{\partial f}{\partial \hat{x}_i} - \frac{\hat{x}_i}{m} \sum_{j=1}^{m} \left(\hat{\chi}_j \cdot \frac{\partial f}{\partial \hat{x}_j}\right)\right] \tag{6}$$

2 Practical

We wanted to compare apples to apples so we tried to keep most hyper-parameters consistent across experiments, for educational purposes (it is clear that if we would do grid search for example most likely that hyper-parameters will vary between experiments). As well, we would like to note that we did no perform hyper-parameter tuning, because the networks achieved sufficient performance out of the box. This is way we also did not use a validation data set.
As we wanted to compare apples to apples we regularized the model only in one layer, after the first fully connected layer. We could have also done differently (another hyper-parameter). We used the following hyper-parameters across all the experiments :

(a) Optimizer - ADAM with default β_1 and β_2

(b) Learning rate - 0.001

(c) Epochs - 10. We would like to note that we set seed for each experiment, so if it was a real model (in production), we can perform early stopping and choose the best score. But than it's better to do have a test score as well (because we are "playing" with the model) in order to get a real estimate about the algorithm accuracy.

Hyper-parameters that are unique for each experiment -

1. BatchNorm with hyper-paramerter affine = False
2. Dropout (with 10
3. Weight decay (for all weights, not only one layer) - $2 \cdot 10^{-4}$
4. No regularization

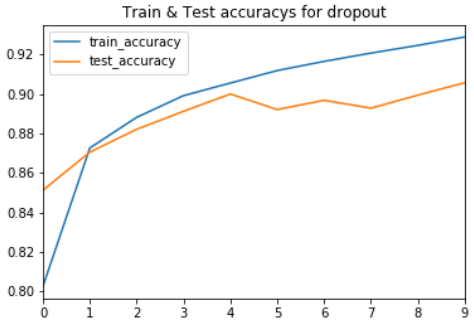
Conclusions

(a) For all the experiments after 2-3 epochs the model over-fits.

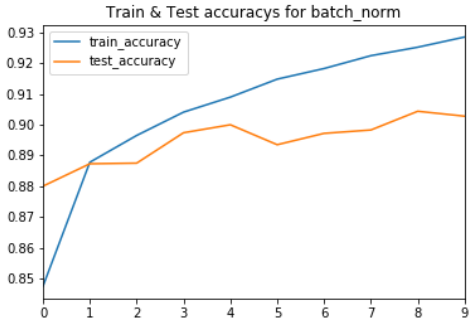
(b) Even though over-fitting happens validation scores keep increasing with the epochs.

(c) All the models are pretty much the same (around 90% accuracy). So in order to argue which model is better it's better to perform an hypothesis test.

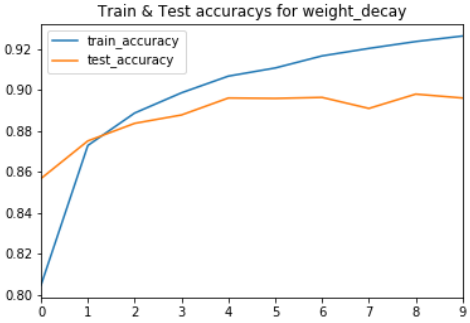
2.1 Graphs



(a)



(b)



(c) 3

(d)

Figure 1: Convergence graphs