Deep Learning for Visual Recognition - Assignment 4

Mayara E. Bonani, Guillaume Rouvarel, Arash Safavi, Vardeep Singh, Cüneyt Erem

December 13, 2020

Theoretical Part

a) Adaptive Learning Rates

1.

$$\frac{\partial L}{\partial w_x} = \frac{\partial}{\partial w_x}\,\sigma(w_x * x + w_y * y)$$

2.

Stochastic gradient descent :

Iteration 1 :

$$\hat{g} = \frac{1}{m}\nabla_\delta \Sigma_i\; L(f(x^{(i)};\delta), y^{(i)}) \tag{1}$$
$$\hat{g} = \frac{1}{m} * \nabla_\theta * \sigma(w_x * x * \theta + w_y * y * \theta) * t \tag{2}$$
$$\hat{g} = \frac{1}{m} * \nabla_\theta * \frac{1}{1 + e^{(w_x * x * \theta + w_y * y * \theta)}} * t \tag{3}$$
$$\hat{g} = \frac{1}{1} * 10^{-8} * \frac{1}{1 + e^{(-1*1*10^{-8} + 1*(-1)*10^{-8})}} * 1 \tag{4}$$
$$\hat{g} = 5.001e^{-9} \tag{5}$$

$$\delta = \delta - \epsilon\hat{g} \tag{6}$$
$$\delta = 10^{-8} - 5.001e^{-9} \tag{7}$$
$$\delta = -4.999e^{-9} \tag{8}$$

Iteration 2 :

$$\hat{g} = \frac{1}{1} * -4.999e^{-9} * \frac{1}{1 + e^{(-1*1*-4.999e^{-9}+1*(-1)*-4.999e^{-9})}} * 1 \tag{9}$$
$$\hat{g} = -2.499e^{-9} \tag{10}$$

$$\delta = -4.999e^{-9} - (-2.499e^{-9}) \tag{11}$$
$$\delta = -2.5e^{-9} \tag{12}$$

Iteration 3 :

$$\hat{g} = \frac{1}{1} * (-2.5e^{-9}) * \frac{1}{1 + e^{(-1*1*(-2.5e^{-9})+1*(-1)*(-2.5e^{-9}))}} * 1 \tag{13}$$
$$\hat{g} = -1.25e{-9} \tag{14}$$

$$\theta = -2.5e^{-9} - (-1.25e{-9}) \tag{15}$$
$$\theta = -1.25e{-9} \tag{16}$$

AdaGrad :
Iteration 1 :

$$g = \frac{1}{m} \nabla_\delta \Sigma_i \, L(f(x^{(i)}; \delta), y^{(i)}) \tag{17}$$
$$g = \frac{1}{m} * \nabla_\theta * \sigma(w_x * x * \theta + w_y * y * \theta) * t \tag{18}$$
$$g = \frac{1}{m} * \nabla_\theta * \frac{1}{1 + e^{(w_x*x*\theta+w_y*y*\theta)}} * t \tag{19}$$
$$g = \frac{1}{1} * 10^{-8} * \frac{1}{1 + e^{(-1*1*10^{-8}+1*(-1)*10^{-8})}} * 1 \tag{20}$$
$$g = 5.001e^{-9} \tag{21}$$

$$r = r + g * g \tag{22}$$
$$r = 0 + 5.001e^{-9} * 5.001e^{-9} \tag{23}$$
$$r = 2.501e^{-17} \tag{24}$$

$$\Delta\theta = -\frac{\epsilon}{\delta + \sqrt{r}} * g \tag{25}$$
$$\Delta\theta = -\frac{1}{10^{-8} + \sqrt{2.501e^{-17}}} * 5.001e^{-9} \tag{26}$$
$$\Delta\theta = -1.554e^{-5} \tag{27}$$

$$\theta = \theta + \Delta\theta \tag{28}$$
$$\theta = 10^{-8} + (-1.554e^{-5}) \tag{29}$$
$$\theta = -1.553e^{-5} \tag{30}$$

Iteration 2 :

$$\hat{g} = \frac{1}{1} * (-1.553e^{-5}) * \frac{1}{1 + e^{(-1*1*(-1.553e^{-5})+1*(-1)*(-1.553e^{-5}))}} * 1 \tag{31}$$
$$\hat{g} = 7.765e^{-6} \tag{32}$$

$$r = 2.501e^{-17} + 7.765e^{-6} * 7.765e^{-6} \tag{33}$$
$$r = 6.03e^{-11} \tag{34}$$

$$\Delta\theta = -\frac{1}{10^{-8} + \sqrt{6.03e^{-11}}} * 7.765e^{-6} \tag{35}$$
$$\Delta\theta = -0.999 \tag{36}$$

$$\theta = -1.553e^{-5} + (-0.999) \tag{37}$$
$$\theta = -0.999 \tag{38}$$

Iteration 3 :

$$\hat{g} = \frac{1}{1} * (-0.999) * \frac{1}{1 + e^{(-1*1*(-0.999)+1*(-1)*(-0.999))}} * 1 \tag{39}$$
$$\hat{g} = 7.367 \tag{40}$$

$$r = 6.03e^{-11} + 7.367 * 7.367 \tag{41}$$
$$r = 54.273 \tag{42}$$

$$\Delta\theta = -\frac{1}{10^{-8} + \sqrt{54.273}} * 7.367 \tag{43}$$
$$\Delta\theta = -0.999 \tag{44}$$

$$\theta = -0.999 + (-0.999) \tag{45}$$
$$\theta = 1.998 \tag{46}$$

RMSProp :
Iteration 1 :

$$g = \frac{1}{m}\nabla_\delta \Sigma_i \; L(f(x^{(i)};\delta), y^{(i)}) \tag{47}$$
$$g = \frac{1}{m} * \nabla_\theta * \sigma(w_x * x * \theta + w_y * y * \theta) * t \tag{48}$$
$$g = \frac{1}{m} * \nabla_\theta * \frac{1}{1 + e^{(w_x * x * \theta + w_y * y * \theta)}} * t \tag{49}$$
$$g = \frac{1}{1} * 10^{-8} * \frac{1}{1 + e^{(-1*1*10^{-8} + 1*(-1)*10^{-8})}} * 1 \tag{50}$$
$$g = 5.001e^{-9} \tag{51}$$

$$r = pr + (1-p) * g * g \tag{52}$$
$$r = 0.9 * 0 + (1-0.9) * 5.001e^{-9} * 5.001e^{-9} \tag{53}$$
$$r = 2.501e^{-18} \tag{54}$$

$$\Delta\theta = -\frac{\epsilon}{\sqrt{\delta + r}} * g \tag{55}$$
$$\Delta\theta = -\frac{1}{\sqrt{10^{-8} + 2.501e^{-18}}} * 5.001e^{-9} \tag{56}$$
$$\Delta\theta = -5.001^{-5} \tag{57}$$

$$\theta = \theta + \Delta\theta \tag{58}$$
$$\theta = 10^{-8} + (-5.001e^{-5}) \tag{59}$$
$$\theta = -5.002^{-5} \tag{60}$$

Iteration 2 :

$$g = \frac{1}{1} * 10^{-8} * \frac{1}{1 + e^{(-1*1*(-5.002^{-5}) + 1*(-1)*(-5.002^{-5}))}} * 1 \tag{61}$$
$$g = 4.999e^{-8} \tag{62}$$

$$r = 0.9 * 2.501e^{-18} + (1 - 0.9) * 4.999e^{-8} * 4.999e^{-8} \tag{63}$$
$$r = 2.522^{-16} \tag{64}$$

$$\Delta\theta = -\frac{1}{\sqrt{10^{-8} + 2.522^{-16}}} * 4.999e^{-8} \tag{65}$$
$$\Delta\theta = -4.999e^{-4} \tag{66}$$

$$\theta = -5.002^{-5} + (-4.999e^{-4}) \tag{67}$$
$$\theta = -5,499e^{-4} \tag{68}$$

Iteration 3 :

$$g = \frac{1}{1} * 10^{-8} * \frac{1}{1 + e^{(-1*1*(-5,499e^{-4})+1*(-1)*(-5,499e^{-4}))}} * 1 \tag{69}$$
$$g = 0.033 \tag{70}$$

$$r = 0.9 * 2.501e^{-18} + (1 - 0.9) * 0.033 * 0.033 \tag{71}$$
$$r = 1.089e^{-4} \tag{72}$$

$$\Delta\theta = -\frac{1}{\sqrt{10^{-8} + 1.089e^{-4}}} * 0.033 \tag{73}$$
$$\Delta\theta = -3.162 \tag{74}$$

$$\theta = -5.002^{-5} + (-3.162) \tag{75}$$
$$\theta = -3.162 \tag{76}$$

Adam :
Iteration 1 :

$$g = \frac{1}{m} \nabla_\delta \Sigma_i \ L(f(x^{(i)}; \delta), y^{(i)}) \tag{77}$$
$$g = \frac{1}{m} * \nabla_\theta * \sigma(w_x * x * \theta + w_y * y * \theta) * t \tag{78}$$
$$g = \frac{1}{m} * \nabla_\theta * \frac{1}{1 + e^{(w_x * x * \theta + w_y * y * \theta)}} * t \tag{79}$$
$$g = \frac{1}{1} * 10^{-8} * \frac{1}{1 + e^{(-1*1*10^{-8}+1*(-1)*10^{-8})}} * 1 \tag{80}$$
$$g = 5.001e^{-9} \tag{81}$$

$$t = t + 1 \tag{82}$$
$$t = 0 + 1 \tag{83}$$
$$t = 1 \tag{84}$$

$$s = p_1 s + (1 - p_1) * g \tag{85}$$
$$s = 0.9 * 0 + (1 - 0.9) * 5.001e^{-9} \tag{86}$$
$$s = 5.001e^{-10} \tag{87}$$

$$r = p_2 r + (1 - p_2) * g * g \tag{88}$$
$$r = 0.999 * 0 + (1 - 0.999) * 5.001e^{-9} * 5.001e^{-9} \tag{89}$$
$$r = 2.501e^{-20} \tag{90}$$

$$\hat{s} = \frac{s}{1 - p_1^t} \tag{91}$$
$$\hat{s} = \frac{5.001e^{-10}}{1 - 0.9} \tag{92}$$
$$\hat{s} = 5.001e^{-9} \tag{93}$$

$$r = \frac{r}{1 - p_2^t} \tag{94}$$
$$r = \frac{2.501e^{-20}}{1 - 0.999} \tag{95}$$
$$r = 2.501e^{-17} \tag{96}$$

$$\Delta\theta = -\epsilon * \frac{\hat{s}}{\sqrt{r^+ \delta}} \tag{97}$$
$$\Delta\theta = -1 * \frac{5.001e^{-9}}{\sqrt{2.501e^{-17} + 10^{-8}}} \tag{98}$$
$$\Delta\theta = -5.001e^{-5} \tag{99}$$

$$\theta = \theta + \Delta\theta \tag{100}$$
$$\theta = 10 + (-5.001e^{-5}) \tag{101}$$
$$\theta = -5.001e^{-5} \tag{102}$$

Iteration 2 :

$$g = \frac{1}{1} * 10^{-8} * \frac{1}{1 + e^{(-1*1*-5.001e^{-5}+1*(-1)*-5.001e^{-5})}} * 1 \tag{103}$$
$$g = 5e^{-9} \tag{104}$$

$$t = 1 + 1 \tag{105}$$
$$t = 2 \tag{106}$$

$$s = 0.9 * 5.001e^{-10} + (1 - 0.9) * 5^{-9} \tag{107}$$
$$s = 9.501e^{-10} \tag{108}$$

$$r = 0.999 * 2.501e^{-20} + (1 - 0.999) * 5e^{-9} * 5e^{-9} \tag{109}$$
$$r = 4.998e^{-20} \tag{110}$$

$$\hat{s} = \frac{9.501e^{-10}}{1 - 0.9} \tag{111}$$
$$\hat{s} = 9.501e^{-9} \tag{112}$$

$$r = \frac{4.998e^{-20}}{1 - 0.999} \tag{113}$$
$$r = 4.998e^{-17} \tag{114}$$

$$\Delta\theta = -1 * \frac{9.501e^{-9}}{\sqrt{4.998e^{-17} + 10^{-8}}} \tag{115}$$
$$\Delta\theta = -9.501e^{-5} \tag{116}$$

$$\theta = -5.001e^{-5} + (-9.501e^{-5}) \tag{117}$$
$$\theta = -1.45e^{-4} \tag{118}$$

Iteration 3 :

$$g = \frac{1}{1} * 10^{-8} * \frac{1}{1 + e^{(-1*1*-1.45e^{-4}+1*(-1)*-1.45e^{-4})}} * 1 \tag{119}$$
$$g = 4.999e^{-9} \tag{120}$$

$$t = 2 + 1 \tag{121}$$
$$t = 3 \tag{122}$$

$$s = 0.9 * 5.001e^{-10} + (1 - 0.9) * 4.999e^{-9} \tag{123}$$
$$s = 9.5e^{-10} \tag{124}$$

$$r = 0.999 * 2.501e^{-20} + (1 - 0.999) * 4.999e^{-9} * 4.999e^{-9} \tag{125}$$
$$r = 4.997e^{-20} \tag{126}$$

$$\hat{s} = \frac{9.5e^{-10}}{1 - 0.9} \tag{127}$$
$$\hat{s} = 9.5e^{-9} \tag{128}$$

$$r = \frac{4.997e^{-20}}{1 - 0.999} \tag{129}$$
$$r = 4.997e^{-17} \tag{130}$$

$$\Delta\theta = -1 * \frac{9.5e^{-9}}{\sqrt{4.997e^{-17} + 10^{-8}}} \tag{131}$$
$$\Delta\theta = -9.5e^{-5} \tag{132}$$

$$\theta = -1.45e^{-4} + (-9.5e^{-5}) \tag{133}$$
$$\theta = -2.4e^{-4} \tag{134}$$

b) Unstable Gradient Problem

b) 1.

$$\begin{aligned}
\frac{\partial h_n}{\partial w_i} &= \frac{\partial \sigma(w_n \cdot h_{n-1})}{\partial (w_n \cdot h_{n-1})} \frac{\partial (w_n \cdot h_{n-1})}{\partial w_i} \\
&= \frac{\partial \sigma(w_n \cdot h_{n-1})}{\partial (w_n \cdot h_{n-1})} \left[\frac{\partial w_n}{\partial w_i} \cdot h_{n-1} + w_n \cdot \frac{\partial h_{n-1}}{\partial w_i} \right] \\
&= \frac{\partial \sigma(w_n \cdot h_{n-1})}{\partial (w_n \cdot h_{n-1})} \left[\delta_{n,i} \cdot h_{n-1} + w_n \cdot \frac{\partial h_{n-1}}{\partial w_i} \right],
\end{aligned}$$

where $\delta_{n,i} = 1$ if $n = i$ or $\delta_{n,i} = 0$ if $n \neq i$. The derivative $\frac{\partial h_{n-1}}{\partial w_i}$ is calculated as before, that is

$$\begin{aligned}
\frac{\partial h_{n-1}}{\partial w_i} &= \frac{\partial \sigma(w_{n-1} \cdot h_{n-2})}{\partial (w_{n-1} \cdot h_{n-2})} \frac{\partial (w_{n-1} \cdot h_{n-2})}{\partial w_i} \\
&= \frac{\partial \sigma(w_{n-1} \cdot h_{n-2})}{\partial (w_{n-1} \cdot h_{n-2})} \left[\frac{\partial w_{n-1}}{\partial w_i} \cdot h_{n-2} + w_{n-1} \cdot \frac{\partial h_{n-2}}{\partial w_i} \right] \\
&= \frac{\partial \sigma(w_{n-1} \cdot h_{n-2})}{\partial (w_{n-1} \cdot h_{n-2})} \left[\delta_{n-1,i} \cdot h_{n-2} + w_{n-1} \cdot \frac{\partial h_{n-2}}{\partial w_i} \right],
\end{aligned}$$

and so on, where where $\delta_{n-1,i} = 1$ if $n-1 = i$ or $\delta_{n-1,i} = 0$ if $n-1 \neq i$. The result above is valid for $i \neq 1$, because $h_1 = \sigma(x)$, thus h_1 and h_n are independent of w_1. Therefore $\frac{\partial h_1}{\partial w_1} = 0$ and

$$\frac{\partial h_n}{\partial w_1} = 0,$$

for any value of n.

b) 2.

For the sigmoid function, the maximum value value of the gradient of weight w_i will be when $i = n$. For instance, for n = 3:

$$\begin{aligned}
\frac{\partial h_3}{\partial w_3} &= \sigma(w_3 h_2)(1 - \sigma(w_3 h_2)) \cdot h_2 \\
\frac{\partial h_3}{\partial w_2} &= \sigma(w_3 h_2) \cdot (1 - \sigma(w_3 h_2)) \cdot w_3 \cdot \sigma(w_2 h_1) \cdot (1 - \sigma(w_2 h_1)) \cdot h_1 \\
\frac{\partial h_3}{\partial w_1} &= 0.
\end{aligned}$$

Since $|w_i| < 1$ and $1 - \sigma(w_i h_{i-1}) < 1$, because the sigmoid is between zero and one, the derivative with respect to w_3 is larger than the one with respect to w_2. Thus the maximum value of the gradient is when $i = n$.

Similarly, for the ReLU activation function, the maximum value value of the gradient of weight w_i will be when $i = n$, if $w_n.h_{n-1}$ is greater than 0. The derivative of the the Relu function is equal to 1 if the argument of the function is greater than 0, and it is 0 if it is smaller than 0.

b) 3.

$\text{Var}(XY) = E(X^2Y^2) - (E(XY))^2 = \text{Var}(X)\text{Var}(Y) + \text{Var}(X)(E(Y))^2 + \text{Var}(Y)(E(X))^2$

Considering $E(\hat{X}) = E(\hat{Y}) = 0$ We obtain:

$$Var(\hat{X}\hat{Y}) = Var(\hat{X})Var(\hat{Y})$$

The same apply for n independent variables which expected value equal to null as it is shown bellow.

$$\begin{aligned}
\text{var}(X_1 \cdots X_n) &= E[(X_1 \cdots X_n)^2] - (E[X_1 \cdots X_n])^2 \\
&= E[X_1^2 \cdots X_n^2] - (E[(X_1] \cdots E[X_n])^2 \\
&= E[X_1^2] \cdots E[X_n^2] - (E[X_1])^2 \cdots (E[X_n])^2 \\
&= \prod_{i=1}^{n} \left(\text{var}(X_i) + (E[X_i])^2\right) - \prod_{i=1}^{n} (E[X_i])^2
\end{aligned}$$

Considering all the expected values equal to zero, we obtain:

$$\text{var}(X_1 \cdots X_n) = \prod_{i=1}^{n} (\text{var}(X_i))$$

b) 4.

We know that

$$h_i^j = \sum_k W_{j,k}^i h^{i-1}$$

Thus the variance for the hidden layer can be expressed as:

$$Var(h_j^i) = Var(\sum_k W_{j,k}^i) Var(h^{i-1})$$

Now since the variance is equal for all the weights. This equation can be expressed as:

$$Var(h_j^i) = n_i Var(W^i) Var(h^{i-1})$$

b) 5.

If there is no change in variance between the layers i.e.

$$Var(h^{i-1}) = Var(h^i)$$

Then the above expression can be simplified to:

$$n_i Var(W^i) = 1$$

$$\therefore Var(W^i) = 1/n_i$$

b) 6.