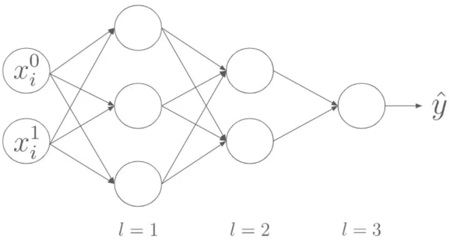
Homework 1

Problem 1

Question 1



setup: $x=\begin{pmatrix}2\\1\end{pmatrix}$ $y=10$ $W^{(1)}=\begin{pmatrix}1&0\\3&4\\1&2\end{pmatrix}$ $b^{(1)}=\begin{pmatrix}0\\-2\\5\end{pmatrix}$ $W^{(2)}=\begin{pmatrix}2&6&1\\0&3&1\end{pmatrix}$ $b^{(2)}=\begin{pmatrix}-23\\-11\end{pmatrix}$

$W^{(3)}=(-2\ 3)$ $b^{(3)}=18$

Analysis: ① $z^{(0)}=x=\begin{pmatrix}2\\1\end{pmatrix}$

② $a^{(1)}=W^{(1)}z^{(0)}+b^{(1)}=\begin{pmatrix}1&0\\3&4\\1&2\end{pmatrix}\begin{pmatrix}2\\1\end{pmatrix}+\begin{pmatrix}0\\-2\\5\end{pmatrix}=\begin{pmatrix}2\\10\\4\end{pmatrix}+\begin{pmatrix}0\\-2\\5\end{pmatrix}=\begin{pmatrix}2\\8\\9\end{pmatrix}$

③ $z^{(1)}=\text{ReLU}(a^{(1)})=\begin{pmatrix}2\\8\\9\end{pmatrix}$

④ $a^{(2)}=W^{(2)}z^{(1)}+b^{(2)}=\begin{pmatrix}2&6&1\\0&3&1\end{pmatrix}\begin{pmatrix}2\\8\\9\end{pmatrix}+\begin{pmatrix}-23\\-11\end{pmatrix}=\begin{pmatrix}61\\33\end{pmatrix}+\begin{pmatrix}-23\\-11\end{pmatrix}=\begin{pmatrix}38\\22\end{pmatrix}$

④ $z^{(2)}=\text{ReLU}(a^{(2)})=\begin{pmatrix}38\\22\end{pmatrix}$

⑤ $a^{(3)}=W^{(3)}z^{(2)}+b^{(3)}=(-2\ 3)\begin{pmatrix}38\\22\end{pmatrix}+18=8$

⑥ $z^{(3)}=\text{ReLU}(a^{(3)})=8$

⑦ $\hat{y}=z^{(3)}=8$

Question 2 $L(\hat{y}, y) = (y-\hat{y})^2$ $\quad ReLU(x) = \max(0,x) \Rightarrow ReLU'(x) = \begin{cases} 0, & x \le 0 \\ 1, & x > 0 \end{cases}$

① $\frac{\partial L}{\partial \hat{y}} = -2(y-\hat{y}) = 2(\hat{y}-y)$

② $\frac{\partial L}{\partial z^{(3)}} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z^{(3)}} = 2(\hat{y}-y)$

③ $\frac{\partial L}{\partial a^{(3)}} = \frac{\partial L}{\partial z^{(3)}} \cdot \frac{\partial z^{(3)}}{\partial a^{(3)}} = 2(\hat{y}-y) \cdot ReLU'(a^{(3)})$

④ $\frac{\partial L}{\partial w^{(3)}} = \frac{\partial L}{\partial a^{(3)}} \cdot \frac{\partial a^{(3)}}{\partial w^{(3)}} = 2(\hat{y}-y) \cdot ReLU'(a^{(3)}) \cdot z^{(2)}$

⑤ $\frac{\partial L}{\partial b^{(3)}} = \frac{\partial L}{\partial a^{(3)}} \cdot \frac{\partial a^{(3)}}{\partial b^{(3)}} = 2(\hat{y}-y) \cdot ReLU'(a^{(3)})$

⑥ $\frac{\partial L}{\partial z^{(2)}} = \frac{\partial L}{\partial a^{(3)}} \frac{\partial a^{(3)}}{\partial z^{(2)}} = 2(\hat{y}-y) \cdot ReLU'(a^{(3)}) \cdot W^{(3)}$

⑦ $\frac{\partial L}{\partial a^{(2)}} = \frac{\partial L}{\partial z^{(2)}} \frac{\partial z^{(2)}}{\partial a^{(2)}} = 2(\hat{y}-y) \cdot ReLU'(a^{(3)}) W^{(3)} ReLU'(a^{(2)})$

⑧ $\frac{\partial L}{\partial w^{(2)}} = \frac{\partial L}{\partial a^{(2)}} \frac{\partial a^{(2)}}{\partial w^{(2)}} = 2(\hat{y}-y) ReLU'(a^{(3)}) W^{(3)} ReLU'(a^{(2)}) \cdot z^{(1)}$

⑨ $\frac{\partial L}{\partial b^{(2)}} = \frac{\partial L}{\partial a^{(2)}} \frac{\partial a^{(2)}}{\partial b^{(2)}} = 2(\hat{y}-y) ReLU'(a^{(3)}) W^{(3)} ReLU'(a^{(2)})$

⑩ $\frac{\partial L}{\partial z^{(1)}} = \frac{\partial L}{\partial a^{(2)}} \frac{\partial a^{(2)}}{\partial z^{(1)}} = 2(\hat{y}-y) ReLU'(a^{(3)}) W^{(3)} ReLU'(a^{(2)}) W^{(2)}$

⑪ $\frac{\partial L}{\partial a^{(1)}} = \frac{\partial L}{\partial z^{(1)}} \frac{\partial z^{(1)}}{\partial a^{(1)}} = 2(\hat{y}-y) ReLU'(a^{(3)}) W^{(3)} ReLU'(a^{(2)}) W^{(2)} ReLU'(a^{(1)})$

⑫ $\frac{\partial L}{\partial w^{(1)}} = \frac{\partial L}{\partial a^{(1)}} \frac{\partial a^{(1)}}{\partial w^{(1)}} = 2(\hat{y}-y) ReLU'(a^{(3)}) w^{(3)} ReLU'(a^{(2)}) w^{(2)} ReLU'(a^{(1)}) z^{(0)}$

⑬ $\frac{\partial L}{\partial b^{(1)}} = \frac{\partial L}{\partial a^{(1)}} \frac{\partial a^{(1)}}{\partial b^{(1)}} = 2(y-y) ReLU'(a^{(3)}) W^{(2)} ReLU'(a^{(2)}) W^{(2)} ReLU'(a^{(1)})$

14 $\frac{\partial L}{\partial z^{(0)}} = \frac{\partial L}{\partial a^{(1)}} \frac{\partial a^{(1)}}{\partial z^{(0)}} = 2(\hat{y}-y) \mathrm{ReLU}'(a^{(3)}) w^{(3)} \mathrm{ReLU}'(a^{(2)}) w^{(2)} \mathrm{ReLU}'(a^{(1)}) w^{(1)}$

15 $\frac{\partial L}{\partial x} = \frac{\partial L}{\partial z^{(0)}} \frac{\partial z^{(0)}}{\partial x} = 2(\hat{y}-y) \mathrm{ReLU}'(a^{(3)}) w^{(3)} \mathrm{ReLU}'(a^{(2)}) w^{(2)} \mathrm{ReLU}'(a^{(1)}) w^{(1)}$

Question 3

① $\ell=3$: $\frac{\partial L}{\partial W^{(3)}} = 2(\hat{y}-y)\cdot ReLU'(a^{(3)})\cdot z^{(2)} = 2\times(8-10)\times 1\times\begin{pmatrix}38\\22\end{pmatrix} = -4\times\begin{pmatrix}38\\22\end{pmatrix} = (-152\ -88)$

$\frac{\partial L}{\partial b^{(3)}} = 2(\hat{y}-y)\cdot ReLU'(a^{(3)}) = 2\times(8-10)\times 1 = -4$

② $\ell=2$: $\frac{\partial L}{\partial W^{(2)}} = 2(\hat{y}-y)ReLU'(a^{(3)})W^{(3)}ReLU'(a^{(2)})\cdot z^{(1)} = 2\times(8-10)\times1\times(-2\ 3)^T\times1\times\begin{pmatrix}2\\8\\9\end{pmatrix}^T$

$= -4\times(-2\ 3)^T(2\ 8\ 9) = \begin{pmatrix}8\\-12\end{pmatrix}(2\ 8\ 9) = \begin{pmatrix}16 & 64 & 72\\-24 & -96 & -108\end{pmatrix}$

$\frac{\partial L}{\partial b^{(2)}} = 2(\hat{y}-y)ReLU'(a^{(3)})W^{(3)}ReLU'(a^{(2)}) = 2\times(8-10)\times1\times(-2\ 3)^T\times1 = \begin{pmatrix}8\\-12\end{pmatrix}$

③ $\ell=1$: $\frac{\partial L}{\partial W^{(1)}} = 2(\hat{y}-y)ReLU'(a^{(3)})W^{(3)}ReLU'(a^{(2)})\,W^{(2)}ReLU'(a^{(1)})\,z^{(0)}$

$=\left(2\times(8-10)\times1\times(-2\ 3)\ \times1\times\begin{pmatrix}2&6&1\\0&3&1\end{pmatrix}\right)^T\times1\times\begin{pmatrix}2\\1\end{pmatrix}^T$

$=\left(-4\times(-2\ 3)\times\begin{pmatrix}2&6&1\\0&3&1\end{pmatrix}\right)^T\times\begin{pmatrix}2\\1\end{pmatrix}^T$

$=\left((8\ -12)\times\begin{pmatrix}2&6&1\\0&3&1\end{pmatrix}\right)^T\times\begin{pmatrix}2\\1\end{pmatrix}^T = \begin{pmatrix}16\\12\\-4\end{pmatrix}(2\ 1) = \begin{pmatrix}32&16\\24&12\\-8&-4\end{pmatrix}$

$\frac{\partial L}{\partial b^{(1)}} = 2(\hat{y}-y)\,ReLU'(a^{(3)})W^{(3)}ReLU'(a^{(2)})\,W^{(2)}ReLU'(a^{(1)})$

$= \begin{pmatrix}16\\12\\-4\end{pmatrix}$

Problem 2

Question 1: number of α: 10, number of λ_2: 6, number of p: 5

Hence, the total number of combination: $10 \times 6 \times 5 = 300$

10 epochs for one combination, then the total epochs are $300 \times 10 = 3000$

And the total time $= 3000$ epoch $\times$ 2min $=$ 6000 min

conclusion: 3000 epochs, 6000 min

Question 2: p(the combination results in good performance) $= 0.05$

P(the combination results in bad performance) $= 1 - 0.05 = 0.95$

P(at least one combination has good results in N trials) $= 1 - 0.95^N$

let $1 - 0.95^N \geq q \Rightarrow 0.95^N \leq 1-q \Rightarrow N \ln 0.95 \leq \ln(1-q) \Rightarrow N \geq \dfrac{\ln(1-q)}{\ln(0.95)}$

① let $q = 0.95$: $N \geq \dfrac{\ln 0.05}{\ln 0.95} \approx 58.4 \Rightarrow 59$ trials

② let $q = 0.995$: $N \geq \dfrac{\ln 0.005}{\ln 0.95} \approx 103.29 \Rightarrow 104$ trials

Problem 3

Question 1: SGD $w_{t+1} = w_t - \alpha \nabla \ell(w_t; x_t)$

① reduce the computational cost

② faster than GD

③ can help the optimizer escape shallow local minima or saddle point

Question 2: Minibatch GD $w_{t+1} = w_t - \alpha \frac{1}{b} \sum_{i=1}^{b} \nabla \ell(w_t, x_i)$

① compared with SGD, it has lower variance

② reduce the computational cost than GD

Question 3: SGD with Momentum $\begin{cases} m_{t+1} = \mu m_t - \alpha \nabla \ell(w_t; x_i) \\ w_{t+1} = w_t + m_{t+1} \end{cases}$

① reduces oscillations and accelerate convergence, especially in problems where gradients vary significantly across dimensions

② enable faster convergence and more stable

Problem 4

Question 1 ① setup: (i) input: 36x36 (ii) filter size: 3x3 (iii) padding: 0 (iv) stripe: 1

② analysis: if the filter starts at column c, then it covers: $c, c+1, c+2$

the last convered column should be: $C+2 \le 36 \Rightarrow C \le 34$

Hence, the starting columns could be 1 through 34, and same logic for row

Then, the output size should be 34 x 34

Question 2: ① setup: (i) input: 36x36x3 (ii) filter: 3x3x3 (iii) padding: 3 (iv) stride: 2 (v) filter size: 2

② analysis: for each channel, it will be calculated by one filter.

(i) If padding size = 3, then the input size will be $\boxed{42}$x42x3 (36+3+3)

same logic as Question 1, suppose c is the starting column, then the filter covers: $C, C+1, C+2$.

Then, $C+2 \le 42 \Rightarrow C \le 40$

☆ (ii) If stripe = 2, then the starting position should be 1, 3, 5, …, 39

From 1 to 39, there will be $1+3+\dots+39 = \underbrace{\frac{39-1}{2}} + 1 = 20$ steps

↳ $1+nd=39 \Rightarrow n = \frac{39-1}{d} = \frac{39-1}{2}$

So there are 20 horizontal positions

(iii) There are 2 filters, then each produces one feature map

so output has 2 channels

Conclusion: 20 x 20 x 2 (height x width x channels)

Question 3: ① set up: (i) input: $n \times n \times c$ (ii) filter size: $k \times k \times c$

(iii) number of filters: ℓ (iv) stride: S (v) padding: P

② analysis:

(i) from padding $= p$, we could know the input should be $(n+2p) \times (n+2p) \times C$

(ii) if we want to get the valid starting position C, then $(C, C+1, \cdots, C+k-1)$

$C + \boxed{k-1} \leq n+2p \Rightarrow C \leq n+2p-k+1$ (starting row/column)

(iii) from stride $= S$, we could analyze the starting position for row/column

it should be: $1, 1+S, 1+2S, \cdots, 1+mS \leq n+2p-k+1$

then $m \leq \frac{1}{S}(n+2p-k)$, so the max step should be: $\lfloor \frac{1}{S}(n+2p-k) \rfloor$

the number of starting points = number of steps $+1 = \lfloor \frac{1}{S}(n+2p-k) \rfloor + 1$

(iv) there are ℓ filters, then there are ℓ feature maps, and the output should have ℓ channels

conclusion: $n_{out} \times n_{out} \times \ell$ and $n_{out} = \lfloor \frac{1}{S}(n+2p-k) \rfloor + 1$

Question4 ① setup: (i) input: $I=\begin{pmatrix} 6 & 0 & 1 \\ 3 & 4 & 1 \\ 1 & 7 & 0 \end{pmatrix}$

(ii) convolutional kernel: $K=\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$

(iii) stripe: 1

(iv) padding: 0

② analysis: $\text{output}=\begin{pmatrix} 6\times1+0\times0+3\times0+4\times(-1) & 0\times1+1\times0+4\times0+1\times(-1) \\ 3\times1+4\times0+1\times0+7\times(-1) & 4\times1+1\times0+7\times0+0\times(-1) \end{pmatrix}$

$=\begin{pmatrix} 2 & -1 \\ -4 & 4 \end{pmatrix}$

Problem 5

Question 1

$$\frac{\partial L}{\partial Z_i} = \frac{\partial L}{\partial Z_m} \frac{\partial Z_m}{\partial Z_{m-1}} \cdots \frac{\partial Z_{i+1}}{\partial Z_i}$$

if $\frac{\partial Z_{i+1}}{\partial Z_i} << 1$, it will make the dot of gradients close to 0. then $\frac{\partial L}{\partial Z_i} \to 0$, which will make model cannot learn weights effectively

Question 2

setup: $\frac{\partial L}{\partial Z_2} = 0.01$, $\frac{\partial Z_2}{\partial Z_1} = 0.005$, $\frac{\partial Z_1}{\partial X} = 0.1$

analysis: $\frac{\partial L}{\partial X} = \frac{\partial L}{\partial Z_2} \times \frac{\partial Z_2}{\partial Z_1} \times \frac{\partial Z_1}{\partial X} = 0.01 \times 0.005 \times 0.1 = 0.000005$

Question 3

setup: $\frac{\partial L}{\partial y} = 0.01$ $\quad y = Z_2 + X \Rightarrow \frac{\partial y}{\partial X} = \frac{\partial Z_2}{\partial X} + 1$

$$\frac{\partial L}{\partial X} = \frac{\partial L}{\partial y} \frac{\partial y}{\partial X} = \frac{\partial L}{\partial y}\left(\frac{\partial Z_2}{\partial X} + 1\right) = \frac{\partial L}{\partial y}\left(\frac{\partial Z_2}{\partial Z_1} \frac{\partial Z_1}{\partial X} + 1\right)$$

$$= 0.01 \times (0.005 \times 0.1 + 1) = 0.010005$$

Question 4

It is obvious that the gradient with residual connection is larger than that without it ($0.010005 > 0.000005$

Question5

1 layer: $\frac{\partial L}{\partial x} = \frac{\partial L}{\partial z_1} \frac{\partial z_1}{\partial x} = 0.5 \times 0.5 = 0.25$

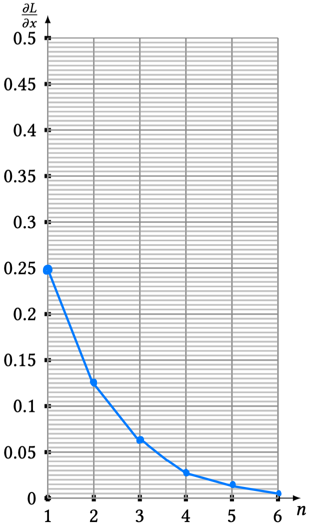
2 layers: $\frac{\partial L}{\partial x} = \frac{\partial L}{\partial z_2} \frac{\partial z_2}{\partial z_1} \frac{\partial z_1}{\partial x} = 0.5^2 \times 0.5 = 0.125$

3 layers: $\frac{\partial L}{\partial x} = \frac{\partial L}{\partial z_3} \frac{\partial z_3}{\partial z_2} \frac{\partial z_2}{\partial z_1} \frac{\partial z_1}{\partial x} = 0.5^4 = 0.0625$

4 layers: $\frac{\partial L}{\partial x} = \frac{\partial L}{\partial z_4} \frac{\partial z_4}{\partial z_3} \frac{\partial z_3}{\partial z_2} \frac{\partial z_2}{\partial z_1} \frac{\partial z_1}{\partial x} = 0.5^6 = 0.03125$

5 layers: $\frac{\partial L}{\partial x} = \frac{\partial L}{\partial z_5} \frac{\partial z_5}{\partial z_4} \frac{\partial z_4}{\partial z_3} \frac{\partial z_3}{\partial z_2} \frac{\partial z_2}{\partial z_1} \frac{\partial z_1}{\partial x} = 0.5^6 = 0.015625$

6 layers: $\frac{\partial L}{\partial x} = \frac{\partial L}{\partial z_6} \frac{\partial z_6}{\partial z_5} \frac{\partial z_5}{\partial z_4} \frac{\partial z_4}{\partial z_3} \frac{\partial z_3}{\partial z_2} \frac{\partial z_2}{\partial z_1} \frac{\partial z_1}{\partial x} = 0.5^7 = 0.0078125$



Question 6

① gradient of a single residual block:

analysis: n conv eayers and each residual block has 2 conv, then the number of residual blocks is n/2

output of j residual block: y_{j-1}

outputs of two conv: z_{2j-1}, z_{2j}

output of block: $y_j = z_{2j} + y_{j-1}$

$$\frac{\partial y_j}{\partial y_{j-1}} = \frac{\partial z_{2j}}{\partial y_{j-1}} + 1 \Rightarrow \frac{\partial z_{2j}}{\partial y_{j-1}} = \frac{\partial z_{2j}}{\partial z_{2j-1}} \frac{\partial z_{2j-1}}{\partial y_{j-1}} = 0.5^2 = 0.25$$

$$\Rightarrow \frac{\partial y_j}{\partial y_{j-1}} = 0.25 + 1 = 1.25$$

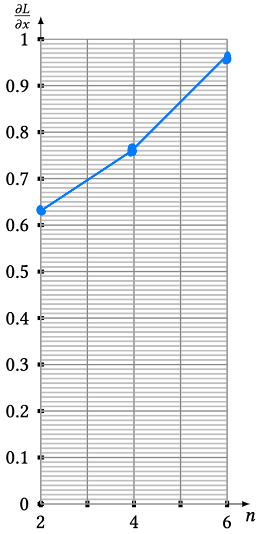
② general: let $g_j = \frac{\partial L}{\partial y_j}$ and $g_B = \frac{\partial L}{\partial y_B} = 05$

$$g_{j-1} = \frac{\partial L}{\partial y_{j-1}} = \frac{\partial L}{\partial y_j} \frac{\partial y_j}{\partial y_{j-1}} = g_j \cdot 1.25 \Rightarrow g_0 = (1.25)^B g_B = 0.5(0.125)^B$$

$n=2 \Rightarrow B=1 \Rightarrow \frac{\partial L}{\partial x} = 0.5 \times 0.125 = 0.625$

$n=4 \Rightarrow B=2 \Rightarrow \frac{\partial L}{\partial x} = 0.5 \times 0.125^2 = 0.78125$

$n=6 \Rightarrow B=3 \Rightarrow \frac{\partial L}{\partial x} = 0.5 \times 0.125^3 = 0.9765625$


$\frac{\partial L}{\partial x}$
n