CSC 420 - Fall 2021 - Assignment 2

Songheng Yin, 1004762303

In solving the questions in this assignment, I worked together with my classmate [Haining Tan, 1004153364]. I confirm that I have written the solutions/code/report in my own words.

Question 1.

For every layer with a linear activation function f. Suppose the matrix representation of f is F and the layer has weights W.

$$F(Wx + b) = FWx + Fb$$

Thus the layer with the activation function is still linear.
For arybitrary neural network with n layers,

$$\begin{aligned}\hat{y} &= f(W_n(f(W_{n-1} \dots f(W_1 x))))) \\ &= (f \circ W_n) \circ (f \circ W_{n-1}) \dots (f \circ W_1)x\end{aligned}$$

Since every component $(f \circ W_i)$ is linear, so the equation can be compressed into $\hat{y} = Wx$.
Therefore there is a nerual network with only one layer which is equivalent to the one of n layers.
That is, the number of layers has no impact.

Question 2.

Denote the three Σ units in the original graph with h_1, h_2, h_3, i.e.
Let

$$\begin{aligned}h_1 &= w_1 x_1 + w_2 x_2 \\ h_2 &= w_3 x_3 + w_4 x_4 \\ h_3 &= w_5 \sigma(h1) + w_6 \sigma(h2)\end{aligned}$$

We also have $\hat{y} = \sigma(h3)$, $L = \|y - \hat{y}\|^2$
Plug in the true value of w, x, we get $h_1 = 1.368, h_2 = -1.432, h_3 = 0.5991, \hat{y} = 0.6454$

Since

$$\frac{\partial L}{\partial \hat{y}} = -2\|y - \hat{y}\|$$
$$\frac{\partial \hat{y}}{\partial h_3} = \sigma(h_3)(1 - \sigma(h_3))$$
$$\frac{\partial h_3}{\partial h_2} = w_6\sigma(h_2)(1 - \sigma(h_2))$$
$$\frac{\partial h_2}{\partial w_3} = x_3$$

By Back Propagation (Chain Rule),

$$\begin{aligned}\frac{\partial L}{\partial w_3} &= \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial h_3}\frac{\partial h_3}{\partial h_2}\frac{\partial h_2}{\partial w_3} \\ &= -2(0.5 - 0.6454)\sigma(0.009)(1 - \sigma(0.009))(-0.2)\sigma(-1.432)(1 - \sigma(-1.432))(-0.3) \\ &= 0.0007\end{aligned}$$

Question 3.

Layer C:
After padding, the input size is $50 \times 14 \times 14$.
Since the stride is 2, for both horizontal and vertical direction, there are $\frac{14-4}{2} + 1 = 6$ positions for the 4×4 filter.
There are 20 filters and 50 channels, hence the total number of resulting matrix elements is $50 \times 20 \times 6 \times 6$.
And for each matrix entry, it comes from doing multiplication 4×4 times, addition $4 \times 4 - 1 = 15$ times, i.e. 31 flops each time.
If we add the bias, then do addition one more time for every output entry.
Put together,

$$50 \times 20 \times 6 \times 6 \times 31 = 1116000 \text{ flops} \qquad \text{Without Bias}$$
$$50 \times 20 \times 6 \times 6 \times (31 + 1) = 1152000 \text{ flops} \qquad \text{With Bias}$$

Layer P:
The pooling layer takes 6×6 size with 20 filters from C as input.
Same as above, due to the 3×3 receptive fields and stride 1, the output size is $\frac{6-3}{1} + 1 = 4$.
The output has shape $20 \times 4 \times 4$.
Also every entry comes from a maximum operation from 3×3 matrix. So,

$$20 \times 4 \times 4 \times (3 \times 3 - 1) = 2560 \text{ flops}$$

In total, the number of flops is 1154560 with bias and 1118560 without bias.

Question 4.

$C1$:
solve: $32 -$ kernel size $+ 1 = 28$, get kernel size is equal to 5.

$$\begin{aligned}\text{number of parameters} &= \text{number of filters} \times \text{number of input channels} \times 5 \times 5 \\ &= 6 \times 5 \times 5 \\ &= 150\end{aligned}$$

$S2$:
Clearly, down-sampling requires 0 trainable parameters.
$C3$:
Same as $C1$, the kernel size is $14 - 10 + 1 = 5$. And

$$\begin{aligned}\text{number of parameters} &= 16 \times 6 \times 5 \times 5 \\ &= 2400\end{aligned}$$

$S4$: 0
$C5$:
It can be treated as a FC from $16 \times 5 \times 5$ to 120. Hence

$$\begin{aligned}\text{number of parameters} &= 16 \times 5 \times 5 \times 120 \\ &= 48000\end{aligned}$$

$F6$, Gaussian Connections: The number of parameters is simply $120 \times 84 + 84 \times 10 = 10920$
The total number of trainable parameters is 61470.

Question 5.

Assume the input of the neural network is vector x, and the first layer has weights w and bias b.
So the output of this layer is $y = \sigma(wx + b)$

$$\begin{aligned}\frac{\partial y}{\partial x} &= \frac{\partial}{\partial x}(\sigma(wx+b)) \\ &= \sigma(wx+b)(1-\sigma(wx+b))w && \text{By Chain Rule} \\ &= \sigma(y)(1-\sigma(y))w && \text{Written in } y\end{aligned}$$

Therefore, we only require y and w to compute the derivative of output with respect to the input. That is, the value of input is not needed.

Question 6.

(a)

When $x \geq 0$, $1 - e^{-2x} \geq 0$. So $\tanh(x) \geq 0$ is increasing because e^{-2x} is decreasing.
When $x < 0$, $1 - e^{-2x} < 0$. So $\tanh(x) < 0$ is increasing because e^{-2x} is decreasing.

Therefore, $\tanh(x)$ is increasing.

$$\lim_{x\to\infty} \tanh(x) = 1$$
$$\lim_{x\to-\infty} \tanh(x) = -1$$

Hence the range of $\tanh(x)$ is $(-1, 1)$.
As learned in class, the output range of logistic function is $(0, 1)$.
A big difference is that $\tanh$ can generate negative outputs.

(b)

$$\begin{aligned}
\tanh'(x) &= \frac{2e^{-2x}(1+e^{-2x}) + 2e^{-2x}(1-e^{-2x})}{(1+e^{-2x})^2} \\
&= \frac{4e^{-2x}}{(1+e^{-2x})^2} \\
&= \frac{4e^{-2x}}{1+e^{-2x}} \frac{1}{1+e^{-2x}} \\
&= \frac{4}{1+e^{2x}} \frac{1}{1+e^{-2x}} \\
&= 4\sigma(-2x)\sigma(2x)
\end{aligned}$$

(c)

Generally speaking, their usage depends on what output range we want.
The output of sigmoid function is always positive, thus in gradient descent, it may lead to zig-zag path sometimes.
Also since the range of sigmoid is $(0, 1)$, we can treat it as the probability in some cases.