Exercise 1.1

Implement NAND gate with a single perceptron

We have definition of NAND:

A	B	A NAND B
0	0	1
0	1	1
1	0	1
1	1	0

And the perceptron:

$$f(x) = \begin{cases} 1, & \text{if } wx+b>0 \\ 0, & \text{otherwise} \end{cases}$$

In our case:

$$f((A,B)) = \begin{cases} 1, & w_1A+w_2B+b>0 \\ 0, & \text{otherwise} \end{cases}$$

We can choose $w=(-0,5;-0,5)$ and $b=1$. Then, when $w\neq(1,1)$ $wx+b>0$ and when $wx=(1,1)$, $wx+b=0$.

Exercise 1.2 Patch-wise DCT transform as convolution.

Describe how the patch-wise DCT transform (with 4×4 patches) of a grayscale image can be implemented and represented using convolutions.

We have DCT_4 for grayscale image:

$$y_k = 2\alpha_k \sum_{j=0}^{N-1} x_j \cdot \cos\left(\pi(j+\frac{1}{2})\frac{k}{N}\right), \quad \alpha_k = \begin{cases} \sqrt{\frac{1}{4N}}, & k=0 \\ \sqrt{\frac{2}{2N}}, & \text{otherwise} \end{cases}$$

In our case, we have patch $4\times4 \Rightarrow N=4$.

We have convolution definition

$$y[k] = (x*z)[k] = \sum_{j=0}^{N-1} x[j]\,z[k-j]$$

Let's perform transformation:

$$y_k = 2\alpha_k \sum_{j=0}^{3} x_j \cdot \cos\left(\pi(j+\frac{1}{2})\frac{k}{4}\right) = \dots = \sum_{j=0}^{3} x_j \cdot \left(2\alpha_k \cos\left(\pi(j+\frac{1}{2})\frac{k}{4}\right)\right)$$

We can say that $x_j = x[j]$'s and second part is $z[k-j]$.
Therefore, $y_k = x * z_k$, $z_k = (z_{kj})$, $z_{kj} = 2_k \cos(\pi(j+\frac{1}{2})\frac{k}{4})$,
$x = (x_j)_{0 \le j \le 3}$, $2_k = \begin{cases} \frac{1}{4}, & k=0 \\ \frac{1}{2\sqrt{2}}, & \text{otherwise} \end{cases}$.

Exercise 1.3

Use the chain rule to compute the gradient of $f(x) = g(Wx+b)$ with respect W, x, b, where g is sigmoid and W is 4×3 matrix.

We have next functions: $f(x) = (g \circ h)(x)$, $h(x) = wx+b$,

$g(y) = \frac{1}{1+e^{-y}}$

$\frac{\partial f}{\partial w} = \frac{\partial f}{\partial h} \cdot \frac{\partial h}{\partial w}$

$\frac{\partial f}{\partial h} = \frac{e^{-h}}{(1+e^{-h})^2}$, $\frac{\partial h}{\partial w} = x^T \Rightarrow \frac{\partial f}{\partial w} = \frac{e^{-(wx+b)}}{(1+e^{-(wx+b)})^2} x^T$

$\frac{\partial f}{\partial b} = \frac{\partial f}{\partial h} \cdot \frac{\partial h}{\partial b}$

$\frac{\partial f}{\partial h} = \frac{e^{-h}}{(1+e^{-h})^2}$, $\frac{\partial h}{\partial w} = 1 \Rightarrow \frac{\partial f}{\partial b} = \frac{e^{-(wx+b)}}{(1+e^{-(wx+b)})^2}$

$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial h} \cdot \frac{\partial h}{\partial x}$

$\frac{\partial f}{\partial h} = \frac{e^{-h}}{(1+e^{-h})^2}$, $\frac{\partial h}{\partial x} = w^T \Rightarrow \frac{\partial f}{\partial x} = \frac{e^{-(wx+b)}}{(1+e^{-(wx+b)})^2} w^T$

Exercise 1.4

Let $f_i(x; \theta_i)$ denote a network layer, with parameter vector θ_i and inputs x. Consider the multi-layer network defined as $F(x) = f_3(y, \theta_3)$, where $y = f_2(f_1(x; \theta_1), \theta_2)$. Also, consider the multi-layer network with skip connection defined as $G(x) = y + f_3(y, \theta_3)$. Use the chain rule to compute the gradient of $F(x)$ with the

respect to all θ_i. Do the same for $G(x)$. What is special in

$\frac{\partial G}{\partial \theta_1} \cdot \frac{\partial F}{\partial \theta_2} = \frac{\partial h_3}{\partial \theta_3}$, $\frac{\partial F}{\partial \theta_2} = \frac{\partial h_3}{\partial h_2} \cdot \frac{\partial h_2}{\partial \theta_2}$, $\frac{\partial F}{\partial \theta_1} = \frac{\partial h_3}{\partial h_2} \cdot \frac{\partial h_2}{\partial h_1} \cdot \frac{\partial h_1}{\partial \theta_1}$

$F(x)$:

$G(x)$: $\frac{\partial G}{\partial \theta_3} = \frac{\partial (y + h_3(y, \theta_3))}{\partial \theta_3} = \frac{\partial y}{\partial \theta_3} + \frac{\partial h_3}{\partial \theta_3} = \frac{\partial h_3}{\partial \theta_3}$, because y is independent from θ_3

$\frac{\partial G}{\partial \theta_2} = \frac{\partial y}{\partial \theta_2} + \frac{\partial h_3}{\partial \theta_2} = \frac{\partial h_2}{\partial \theta_2} + \frac{\partial h_3}{\partial h_2} \cdot \frac{\partial h_2}{\partial \theta_2} =$

$= \frac{\partial h_2}{\partial \theta_2}\left(1 + \frac{\partial h_3}{\partial h_2}\right)$

$\frac{\partial G}{\partial \theta_1} = \frac{\partial y}{\partial \theta_1} + \frac{\partial h_3}{\partial \theta_1} = \frac{\partial h_2}{\partial \theta_1} + \frac{\partial h_3}{\partial \theta_1} = \frac{\partial h_2}{\partial h_1} \cdot \frac{\partial h_1}{\partial \theta_1} + \frac{\partial h_3}{\partial h_2} \cdot \frac{\partial h_2}{\partial h_1} \cdot \frac{\partial h_1}{\partial \theta_1}$

$= \left(\frac{\partial h_2}{\partial h_1} \cdot \frac{\partial h_1}{\partial \theta_1}\right)\left(1 + \frac{\partial h_3}{\partial h_2}\right)$.

The closer ~~output~~ calculations to the end, the ~~vanishing~~ vanishing gradient problem. In $\frac{\partial G}{\partial \theta_1}$ we have $\frac{\partial h_3}{\partial h_2}$ which at some point can approach 0. However, $\frac{\partial G}{\partial \theta_2}$ won't approach 0 in that case because we have $\left(1 + \frac{\partial h_3}{\partial h_2}\right)$.

Therefore, $\frac{\partial G}{\partial \theta_1} \not\to 0$ as $\frac{\partial h_3}{\partial h_2} \to 0$, but $\frac{\partial G}{\partial \theta_1} \to \frac{\partial h_2}{\partial h_1} \cdot \frac{\partial h_1}{\partial \theta_1}$, which is ~~what~~ the feature of skip connections: attenuate vanishing gradient problem.