IMDEN   le 3/11

Homework 5

Exercice 1:

The NAND gate behaves as follow

Input		Output
x_1	x_2	y
0	0	1
1	0	1
0	1	1
1	1	0

let (x_1, x_2) be the input to a perceptron

w_1, w_2, b the weights and bias of the perceptron.

Then $y = f(x_1, x_2) = \begin{cases} 1 & \text{if } w_1 x_1 + w_2 x_2 + b \geq 0 \\ 0 & \text{otherwise} \end{cases}$

is the output of the perceptron.

In that case, we want:

- $f(0,0) = f(1,0) = f(0,1) = 1$
- $f(1,1) = 0$

We can choose

$$\boxed{w_1 = w_2 = -2 \qquad b = 3}.$$

Exercice 1.2:

let x the input image (grayscale)

K the kernel of a convolution layer, of size n.

y the output of $x * K$.

We apply the convolution in 1D.

$$y[k] = \sum_{j=-n} x[j] K[k-j].$$

Let's write a DCT_{N-1} transform in 1D as

$$Y_k = 2\alpha_k \sum_{j=0} X_j \cos\left[\pi\left(j+\frac{1}{2}\right)\frac{k}{N}\right]$$

with $N = 4$ on a 4×4 patch, in 1D

and $k \leq N$

Thus, we can create N kernels: K_k , $k \leq N$
such that
$$Y_k = 2\alpha_k \sum_{j=0}^{N-1} X_j \cos\left[\pi\left(j+\frac{1}{2}\right)\frac{k}{N}\right]$$
$$= \sum_{j=0}^{N-1} X_j \underbrace{2\alpha_k \cos\left[\pi\left(-(k-j)+\overbrace{k}^{+k-k}+\frac{1}{2}\right)\frac{k}{N}\right]}$$
ie $Y[k] = \sum_{j=0}^{N-1} X[j]\, K_k[k-j]$

$$\boxed{\begin{array}{l}\text{Thus, we can set } N \text{ filters with the } k\text{-th filter:} \\ K_k[i] = 2\alpha_k \cos\left[\pi\left(-i+k+\frac{1}{2}\right)\frac{k}{N}\right] \\ \text{and} \quad 0 \leq i \leq N-1.\end{array}}$$

In the end, we have 4 filters of size 4 each.
The DCT transform has been represented as a convolution in 1D
If we do the same process in 2D, we represent DCT as 2 convolutions.

Exercise 13
let $x \in \mathbb{R}^3$
$W \in \mathbb{R}^{4\times 3} = \begin{bmatrix} w_1^T \\ \vdots \\ w_4^T \end{bmatrix}$
$b \in \mathbb{R}^4 = \begin{bmatrix} b_1 \\ \vdots \\ b_4 \end{bmatrix}$
$y = f(x) = g(Wx+b) \in \mathbb{R}^4$
g the sigmoid function defined as $g: u \longmapsto \frac{1}{1+e^{-u}}$

Then, let's compute the gradient of f wrt W, b, x.
Preliminary result:
let's call $\boxed{\tilde{y} = Wx+b}$. Then, for $h_j = \begin{bmatrix} \vdots \\ 1 \\ \vdots \end{bmatrix} \in \mathbb{R}^3$ in j-th
- $(w_i + h_j)^T x - w_i^T x = h_j^T x = x^T h_j$
  So $\frac{d\tilde{y}}{dW_{ij}} = x^T$
- Similarly,
  $\frac{d\tilde{y}}{db_i} = 1$

- $\dfrac{\partial \tilde{y}}{\partial x_i} = W^T$

And

$$\frac{\partial y}{\partial \tilde{y}} = \frac{e^{-\tilde{y}}}{(1+e^{-\tilde{y}})^2} = \frac{1}{e^{\tilde{y}}+2+e^{-\tilde{y}}}$$

And finally, using the chain rule:

- $\dfrac{\partial f}{\partial W} = \dfrac{\partial f}{\partial \tilde{y}} \dfrac{\partial \tilde{y}}{\partial W} = \dfrac{x}{2+e^{-\tilde{y}}+e^{+\tilde{y}}}$

- $\dfrac{\partial f}{\partial x} = \dfrac{\partial f}{\partial \tilde{y}} \dfrac{\partial \tilde{y}}{\partial x} = \dfrac{W^T}{2+e^{-\tilde{y}}+e^{\tilde{y}}}$

- $\dfrac{\partial f}{\partial b} = \dfrac{\partial f}{\partial \tilde{y}} \dfrac{\partial \tilde{y}}{\partial b} = \dfrac{1}{2+e^{-\tilde{y}}+e^{\tilde{y}}}$

To be more precise, we have

- $\boxed{\dfrac{\partial f_k}{\partial W_{ij}} = \dfrac{\partial f_k}{\partial \tilde{y}_k} \times \dfrac{\partial \tilde{y}_k}{\partial W_{ij}} = \begin{cases} 0 \text{ si } k \neq i \\ \dfrac{x_j}{2+e^{-\tilde{y}_k}+e^{\tilde{y}_k}} \text{ sinon.} \end{cases}}$

- $\dfrac{\partial f_k}{\partial x_i} = \dfrac{\partial f_k}{\partial \tilde{y}_k} \times \dfrac{\partial \tilde{y}_k}{\partial x_i}$

- $\boxed{\dfrac{\partial f_k}{\partial x_i} = \dfrac{W_{ki}}{2+e^{-\tilde{y}_k}+e^{+\tilde{y}_k}}}$

- $\boxed{\dfrac{\partial f_k}{\partial b_i} = \dfrac{1}{2+e^{-\tilde{y}_k}+e^{\tilde{y}_k}}}$

Exercice 1.4:

Let $f_i(x, \theta_i)$ a network layer

$F(x)$ the network of 3 f_i layers

$G(x)$ the network of 3 f_i layers + a skip connection in last layer.

Let's compute the gradient, using the chain rule:

We set $\tilde{y} = f_1(x, \theta_1)$

Thus

$$\boxed{\begin{aligned}
&\bullet \frac{\partial F}{\partial \theta_1} = \frac{\partial f_3(y, \theta_3)}{\partial y} \times \frac{\partial f_2(\tilde{y}, \theta_2)}{\partial \tilde{y}} \times \frac{\partial f_1(x, \theta_1)}{\partial \theta_1} \\
&\bullet \frac{\partial F}{\partial \theta_2} = \frac{\partial f_3(y, \theta_3)}{\partial y} \times \frac{\partial f_2(\tilde{y}, \theta_2)}{\partial \theta_2} \\
&\bullet \frac{\partial F}{\partial \theta_3} = \frac{\partial f_3(y, \theta_3)}{\partial \theta_3}
\end{aligned}}$$

If we do the same on G:

$$\bullet \frac{\partial G}{\partial \theta_1} = \left[1 + \frac{\partial f_3(y, \theta_3)}{\partial y}\right] \times \frac{\partial f_2(\tilde{y}, \theta_2)}{\partial \tilde{y}_1} \times \frac{\partial f_1(x, \theta_1)}{\partial \theta_1}$$

ie $$\boxed{\frac{\partial G}{\partial \theta_1} = \frac{\partial F}{\partial \theta_1} + \frac{\partial f_2(\tilde{y}, \theta_2)}{\partial \tilde{y}_1} \times \frac{\partial f_1(x, \theta_1)}{\partial \theta_1}}$$

$$\bullet \boxed{\frac{\partial G}{\partial \theta_2} = \frac{\partial F}{\partial \theta_2} + \frac{\partial f_2(\tilde{y}, \theta_2)}{\partial \theta_2}}$$

$$\bullet \boxed{\frac{\partial G}{\partial \theta_3} = \frac{\partial F}{\partial \theta_3}}$$

Adding a skip connection prevents vanishing gradients as the gradient becomes a sum of multiple terms.