Image Denoising : Course 5

Maxence Gollier

November 2023

1 Exercices

1.1 Exercise 1.1

Let g be the Heaviside function, a NAND gate can be modeled as

$$f(x_1, x_2) = g(1.5 - x_1 - x_2).$$

We verify that we have

$$\begin{aligned} f(0,0) &= g(1.5) = 1, \\ f(0,1) &= g(0.5) = 1, \\ f(1,0) &= g(0.5) = 1, \\ f(1,1) &= g(-0.5) = 0. \end{aligned}$$

By definition, f defines a NAND gate and is a single perceptron with Heaviside activation function.

1.2 Exercise 1.2

The definition of the DCT of a signal X of size N is

$$Y_k = \sum_{j=0}^{N-1} X_j 2\alpha_k \cos(\pi(j + \frac{1}{2})\frac{k}{N}).$$

Rewriting this as a convolution:

$$Y_k = \sum_{j=0}^{N-1} X_j 2\alpha_k \cos(\pi(k + (k-j) + \frac{1}{2})\frac{k}{N}) = X * C_k,$$

where

$$C_k(j) = 2\alpha_k \cos(\pi(k + j + \frac{1}{2})\frac{k}{N})$$

is the convolution kernel. Since we have 4×4 patches we have $N = 16$ and the kernel C is 16×16.

1.3 Exercise 1.3

Let $h := Wx + b$, then we have

$$f = \sigma(h).$$

The chain rule implies that

$$\frac{\partial f}{\partial x} = (\frac{d\sigma}{dh})^T \frac{\partial h}{\partial x},$$
$$\frac{\partial f}{\partial W} = (\frac{d\sigma}{dh})^T \frac{\partial h}{\partial W},$$
$$\frac{\partial f}{\partial b} = (\frac{d\sigma}{dh})^T \frac{\partial h}{\partial b}.$$

We compute

$$\frac{d\sigma}{dh} = \frac{e^{-h}}{(1+e^{-h})^2},$$
$$\frac{\partial h}{\partial x} = W,$$
$$\frac{\partial h}{\partial W} = \begin{pmatrix} x & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & x \end{pmatrix} \in \mathbb{R}^{4\times 4\times 3},$$
$$\frac{\partial h}{\partial b} = I_4.$$

Note : the sigmoid function is computed component-wise since $h \in \mathbb{R}^4$ Hence, we suppose $\frac{d\sigma}{dh} \in \mathbb{R}^4$. Hence, we have

$$\frac{\partial f}{\partial x} = (\frac{e^{-h}}{(1+e^{-h})^2})^T W,$$
$$\frac{\partial f}{\partial W} = (\frac{e^{-h}}{(1+e^{-h})^2})^T \begin{pmatrix} x \\ \vdots \\ x \end{pmatrix},$$
$$\frac{\partial f}{\partial b} = (\frac{e^{-h}}{(1+e^{-h})^2})^T.$$

1.4 Exercise 1.4

Using the chain rule, and defining $z := f_1(x;\theta_1)$ we have

$$\frac{\partial \mathcal{F}}{\partial \theta_3} = \frac{\partial f_3}{\partial \theta_3}, \qquad \frac{\partial \mathcal{G}}{\partial \theta_3} = \frac{\partial f_3}{\partial \theta_3},$$
$$\frac{\partial \mathcal{F}}{\partial \theta_2} = \frac{\partial f_3}{\partial y} \times \frac{\partial f_2}{\partial \theta_2}, \qquad \frac{\partial \mathcal{G}}{\partial \theta_2} = \frac{\partial \mathcal{F}}{\partial \theta_2} + \frac{\partial f_2}{\partial \theta_2},$$
$$\frac{\partial \mathcal{F}}{\partial \theta_1} = \frac{\partial f_3}{\partial y} \times \frac{\partial f_2}{\partial z} \times \frac{\partial f_1}{\partial \theta_1}, \qquad \frac{\partial \mathcal{G}}{\partial \theta_1} = \frac{\partial \mathcal{F}}{\partial \theta_1} + \frac{\partial f_2}{\partial z} \times \frac{\partial f_1}{\partial \theta_1}.$$

We see that adding a skipping term allows to prevent vanishing gradient issues in the last layer f_3: The gradient of $\mathcal{G}(x)$ is the addition of the gradient of the network without skip-connection, which may or may not have vanishing gradient issues and a gradient only depending on the first two layers f_1 and f_2.