Model reference

Saturday, November 9, 2024 6:03 PM

DEFINITIONS:

L = the cross entropy loss function.
σ = the activation function of a neuron.
y = the expected output of the network.
t = the digit associated with the network's input.
a = the activation vector of this layer.
x = the activation vector of the previous layer.
W = the weight matrix of this layer
b = the bias vector of this layer.

$$L = -\sum_k y_k \log a_k = -\log a_t$$
$$z = Wx + b \Longleftrightarrow z_j = W_{jk} x_k + b_j$$
$$a = \sigma(z) = \sigma(Wx + b)$$

For the output layer:

$$\frac{\partial L}{\partial a_j} = \frac{\partial}{\partial a_j}(-\log a_t) = \begin{cases} -1/a_j & \text{if } j = t \\ 0 & \text{if } j \neq t \end{cases}$$

$$\frac{\partial a_j}{\partial z_i} = a_j(\delta_{ij} - a_i)$$

$$\frac{\partial L}{\partial z_i} = \sum_j \frac{\partial L}{\partial a_j} \cdot \frac{\partial a_j}{\partial z_i} = \frac{\partial L}{\partial a_t} \cdot \frac{\partial a_t}{\partial z_i} = -\frac{1}{\cancel{a_t}} \cdot \cancel{a_t}(\delta_{ti} - a_i)$$
$$= -(\delta_{ti} - a_i) = a_i - \delta_{ti} = a_i - y_i$$

$$\rightarrow \boxed{\frac{\partial L}{\partial z} = a - y}$$

$$\frac{\partial z_i}{\partial w_{jk}} = \begin{cases} x_k & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$
$$\frac{\partial L}{\partial w_{jk}} = \sum_i \frac{\partial L}{\partial z_i} \cdot \frac{\partial z_i}{\partial w_{jk}} = \frac{\partial L}{\partial z_j} \cdot \frac{\partial z_j}{\partial w_{jk}} = \frac{\partial L}{\partial z_j} \cdot x_k$$
$$\rightarrow \boxed{\frac{\partial L}{\partial W} = \frac{\partial L}{\partial z} \cdot x^T}$$

$$\frac{\partial z_i}{\partial b_j} = \delta_{ij}$$
$$\frac{\partial L}{\partial b_j} = \sum_i \frac{\partial L}{\partial z_i} \cdot \frac{\partial z_i}{\partial b_j} = \frac{\partial L}{\partial z_j} \cdot \frac{\partial z_j}{\partial b_j} = \frac{\partial L}{\partial z_j}$$
$$\rightarrow \boxed{\frac{\partial L}{\partial b} = \frac{\partial L}{\partial z}}$$

$$\frac{\partial z_i}{\partial x_k} = w_{ik}$$
$$\frac{\partial L}{\partial x_k} = \sum_i \frac{\partial L}{\partial z_i} \cdot \frac{\partial z_i}{\partial x_k} = \sum_i \frac{\partial L}{\partial z_i} \cdot w_{ik}$$
$$\rightarrow \boxed{\frac{\partial L}{\partial x} = W^T \cdot \frac{\partial L}{\partial z}}$$

For other layers

$$\frac{\partial a_j}{\partial z_j} = \mathrm{ReLU}'(z_j) = \begin{cases} 1 & \text{if } z_j > 0 \\ 0 & \text{if } z_j \leq 0 \end{cases} = \begin{cases} 1 & \text{if } a_j > 0 \\ 0 & \text{if } a_j \leq 0 \end{cases}$$
$$\frac{\partial L}{\partial z_j} = \frac{\partial L}{\partial a_j} \cdot \frac{\partial a_j}{\partial z_j} = \frac{\partial L}{\partial a_j} \cdot \mathrm{ReLU}'(z_j)$$
$$\rightarrow \boxed{\frac{\partial L}{\partial z} = \frac{\partial L}{\partial a} \odot \mathrm{ReLU}'(z)}$$

$$\frac{\partial L}{\partial W} = \frac{\partial L}{\partial z} \cdot x^T \qquad \frac{\partial L}{\partial b} = \frac{\partial L}{\partial z} \qquad \frac{\partial L}{\partial x} = W^T \cdot \frac{\partial L}{\partial z}$$