Deduction of Gradients in Neutral Network

Qianyue Hao

November 2020

We use the following letters to represent some intermediate variables:

$$L = -log(q_m) + \frac{\alpha}{2}||W||^2 \tag{1}$$

$$q = softmax(c) \tag{2}$$

$$c = hW_2 + b_2 \tag{3}$$

$$h = ReLU(D) \tag{4}$$

$$D = xW_1 + b_1 \tag{5}$$

$$\begin{aligned}
&\frac{\partial L}{\partial W_1} = \frac{1}{N}(x^T \frac{\partial L}{\partial D}) + \alpha||W_1|| \\
&= \frac{1}{N}\{x^T \mathbb{1}[D \geqslant 0] \circ \frac{\partial L}{\partial h}\} + \alpha||W_1|| \\
&= \frac{1}{N}\{x^T \mathbb{1}[D \geqslant 0] \circ (\frac{\partial L}{\partial c} W_2^T)\} + \alpha||W_1|| \\
&= \frac{1}{N}\{x^T \mathbb{1}[D \geqslant 0] \circ [\frac{\partial L}{\partial q} \circ q_m(\delta_m - q)W_2^T]\} + \alpha||W_1|| \\
&= \frac{1}{N}\{x^T \mathbb{1}[D \geqslant 0] \circ [-\frac{1}{q_m} q_m(\delta_m - q)W_2^T]\} + \alpha||W_1|| \\
&= \frac{1}{N}\{x^T \mathbb{1}[xW_1 + b_1 \geqslant 0] \circ [(q - \delta_m)W_2^T]\} + \alpha||W_1||
\end{aligned} \tag{6}$$

$$\begin{aligned}
&\frac{\partial L}{\partial b_1} = \frac{1}{N}\frac{\partial L}{\partial D} \\
&= \frac{1}{N}\{\mathbb{1}[D \geqslant 0] \circ \frac{\partial L}{\partial h}\} \\
&= \frac{1}{N}\{\mathbb{1}[D \geqslant 0] \circ (\frac{\partial L}{\partial c} W_2^T)\} \\
&= \frac{1}{N}\{\mathbb{1}[D \geqslant 0] \circ [\frac{\partial L}{\partial q} \circ q_m(\delta_m - q)W_2^T]\} \\
&= \frac{1}{N}\{\mathbb{1}[D \geqslant 0] \circ [-\frac{1}{q_m} q_m(\delta_m - q)W_2^T]\} \\
&= \frac{1}{N}\{\mathbb{1}[xW_1 + b_1 \geqslant 0] \circ [(q - \delta_m)W_2^T]\}
\end{aligned} \tag{7}$$

$$
\begin{aligned}
\frac{\partial L}{\partial W_2} &= \frac{1}{N}(h^T\frac{\partial L}{\partial c}) + \alpha||W_2|| \\
&= \frac{1}{N}[h^T\frac{\partial L}{\partial q} \circ q_m(\delta_m - q)] + \alpha||W_2|| \\
&= \frac{1}{N}\{h^T[-\frac{1}{q_m}q_m(\delta_m - q)]\} + \alpha||W_2|| \\
&= \frac{1}{N}[h^T(q - \delta_m)] + \alpha||W_2||
\end{aligned} \tag{8}
$$

$$
\begin{aligned}
\frac{\partial L}{\partial b_2} &= \frac{1}{N}\frac{\partial L}{\partial c} \\
&= \frac{1}{N}[\frac{\partial L}{\partial q} \circ q_m(\delta_m - q)] \\
&= \frac{1}{N}[-\frac{1}{q_m}q_m(\delta_m - q)] \\
&= \frac{1}{N}(q - \delta_m)
\end{aligned} \tag{9}
$$

In the above deduction, N refers batch size; m refers the true class a sample belongs to; q_m refers the m-th element in vector q; δ_m refers the one-hot vector where only the m-th element is 1; $\mathbb{1}[\ ]$ is indicative function.