Mathematical Foundation of DNN : HW 5

Jeong Min Lee

April 2, 2024

1

In this problem, I dervied the representation of each gradient, considering the result of problem6 at hw4. Considering the forward pass in the first half of the code, the following relation is held, which does not agree to the notation in this problem.

$$y[l] = S(A_{list}[l-1]@y[l-1] + b_{list}[l-1])$$

Furthermore, due to the confusion from the notation, I introduced following definitions.

$$db_l = \frac{d}{db_l}loss, \quad dy_l = \frac{d}{dy_l}loss, \quad dA_l = \frac{d}{dA_l}loss$$

According to those notations, the code that calculate the each gradient can be implemented, easily.

$$\begin{aligned} db_l &= \frac{d}{db_l}loss \\ &= \frac{d\ loss}{dy_{l+1}}\frac{dy_{l+1}}{db_l} \\ &= dy_{l+1}\frac{d}{db_l}S(A_l y_l + b_l) \\ &= dy_{l+1}\mathrm{diag}(S'(A_l y_l + b_l)) \end{aligned}$$

$$\begin{aligned} dA_l &= \frac{d}{dA_l}loss \\ &= \frac{d\ loss}{dy_{l+1}}\frac{dy_{l+1}}{dA_l} \\ &= dy_{l+1}\frac{d}{dA_l}S(A_l y_l + b_l) \\ &= y_l dy_{l+1}\mathrm{diag}(S'(A_l y_l + b_l)) \\ &= y_l db_l \end{aligned}$$

$$\begin{aligned} dy_l &= \frac{d}{dy_l}loss \\ &= \frac{d\ loss}{dy_{l+1}}\frac{dy_{l+1}}{dy_l} \\ &= dy_{l+1}\frac{d}{dy_l}S(A_l y_l + b_l) \\ &= dy_{l+1}\mathrm{diag}(S'(A_l y_l + b_l))A_l \\ &= db_l A_l \end{aligned}$$

2

This problem can be solved by applying the chain rule, repeatedly.

$$\begin{aligned} \frac{\partial y_L}{\partial b_i} &= \frac{\partial y_L}{\partial y_{L-1}}\frac{\partial y_{L-1}}{\partial y_{L-2}}\cdots\frac{\partial y_{i+1}}{\partial y_i}\frac{\partial y_i}{\partial b_i} && (1) \\ &= \mathrm{diag}(\tilde{y_L})A_L\mathrm{diag}(\sigma'(\tilde{y}_{L-1}))A_{L-1}\cdots\mathrm{diag}(\sigma'(\tilde{y}_{i+1}))A_{i+1}\mathrm{diag}(\sigma'(\tilde{y}_i)) && (2) \end{aligned}$$

$$\frac{\partial y_L}{\partial A_i} = \mathrm{diag}(\sigma'(\tilde{y}_i))\left(\frac{\partial y_L}{\partial y_i}\right)^T y_{i-1}^T \qquad (3)$$

where $\frac{\partial y_L}{\partial y_i} = \frac{\partial y_L}{\partial y_{L-1}}\frac{\partial y_{L-1}}{\partial y_{L-2}}\cdots\frac{\partial y_{i+1}}{\partial y_i} = \mathrm{diag}(\tilde{y}_L)A_L\mathrm{diag}(\sigma'(\tilde{y}_{L-1}))A_{L-1}\cdots\mathrm{diag}(\sigma'(\tilde{y}_{i+1}))A_{i+1}$.

As A_i is small, $\frac{\partial y_L}{\partial b_i}$, $\frac{\partial y_L}{\partial y_i}$ become extremely small, since $\mathrm{diag}(\tilde{y}_i)$'s elements are moderate and A_i keep being multiplied.