CS224N PA4: Neural Networks for Named Entity Recognition

Daoying Lin (SUID 06090664) Patrick Manion (SUID)

1 Gradients for Backpropagation

Let w_{jk}^l denote the weight for connecting the k^{th} neuron in the $(l-1)^{th}$ layer to the j^{th} neuron in the l^{th} layer; b_j^l denote the bias for the j^{th} neuron in the l^{th} layer; a_j^l denote the activation of the j^{th} neuron in the l^{th} layer; z_j^l denote the weighted input to the j^{th} neuron in the l^{th} layer; $h_l(.)$ denote the activation function for the weighted input $\mathbf{z}_l$. Note that $z_j^l = \sum_i w_{ji}^l a_i^{l-1} + b_j^l$ and $a_j^l = h_l(z_j^l)$. Let's define $\delta_j^l = \frac{\partial J}{\partial z_j^l}$, the error of neuron j in layer l. Then it can be easily derived that the following four equations are true for any backpropagation system:

$$\delta^L = \frac{\partial J}{\partial a^L} \odot h_L'(z^L) \tag{1a}$$

$$\delta^l = (W^{l+1})^T \delta^{l+1} \odot h_l'(z^l) \tag{1b}$$

$$\frac{\partial J}{\partial b_j^l} = \delta_j^l \tag{1c}$$

$$\frac{\partial J}{\partial w_{jk}^l} = a_k^{l-1} \delta_j^l \tag{1d}$$

For current system, we've three layers: input layer, hidden layer and output layer. The cost function is $J = -[y ln a^L + (1-y) ln(1-a^L)]$. Using the above general system, we can obtain the following:

$$\delta^3 = p_\theta - y \tag{2a}$$

$$\delta^2 = (W^3)^T \delta^3 \odot tanh'(z^2) = U^T (p_\theta - y) \odot tanh'(Wx + b^{(1)}) \tag{2b}$$

$$\delta^1 = (W^2)^T \delta^2 \odot I'(x) = W^T \delta^2 = W^T U^T \delta^3 \odot tanh'(Wx + b^{(1)}) \tag{2c}$$

And

$$\frac{\partial J}{\partial U} = a^2\delta^3 = tanh(Wx + b^{(1)})(p_\theta - y) \tag{3a}$$

$$\frac{\partial J}{\partial W} = a^1\delta^2 = LU^T(p_\theta - y) \odot tanh'(Wx + b^{(1)}) \tag{3b}$$

$$\frac{\partial J}{\partial L} = a^0\delta^1 = W^T U^T(p_\theta - y) \odot tanh'(Wx + b^{(1)}) \tag{3c}$$

$$\frac{\partial J}{\partial b^{(2)}} = \delta^3 = p_\theta - y \tag{3d}$$

$$\frac{\partial J}{\partial b^{(1)}} = \delta^2 = U^T(p_\theta - y) \odot tanh'(Wx + b^{(1)}) \tag{3e}$$