Derivation of gradients for the RNTN

Hjalmar K. Turesson

July 24, 2015

Notation

- d – Length of word vector
- n – Node/layer
- x – Activation/output of neuron ($x \in \mathbb{R}^d$; $\tanh z$
- z – Input to neuron ($z \in \mathbb{R}^d$; $z = Wx$)
- t – Target vector ($t \in \mathbb{R}^5$; 0-1 coded)
- y – Prediction ($y \in \mathbb{R}^5$; output of softmax layer – $softmax(z)$)
- W_s – Classification matrix ($W_s \in \mathbb{R}^{5\times d}$)
- W – Weight matrix ($W \in \mathbb{R}^{d\times 2d}$)
- V – Weight tensor ($V^{1:d} \in \mathbb{R}^{2d\times 2d\times d}$)
- L – Word embedding matrix ($L \in \mathbb{R}^{d\times |V|}$, $|V|$ is the size of the vocabulary)
- θ – All weight parameters ($\theta = (W_s, W, V, L)$)
- E – The cost as a function of θ
- δ_l – Error going to the left child node (δ_r error to the right child node)

Softmax

$$y_i = \frac{e^{z_i}}{\sum\limits_j e^{z_j}} \tag{1}$$

$$\frac{\partial y_i}{\partial z_j} = y_i(\delta_{ij} - y_j) \tag{2}$$

δ_{ij} is the Kronecker's delta: $\delta_{ij} = \begin{cases} 0 & \text{if } i \neq j, \\ 1 & \text{if } i = j. \end{cases}$

Cost function E

$$E(\theta) = -\sum_i \sum_j t_j^i \log y_j^i + \lambda ||\theta||^2 \tag{3}$$

$$\frac{\partial E}{\partial y_j} = \frac{t_j}{y_j} \tag{4}$$

Activation function

$$x_i = \tanh z_i \tag{5}$$

$$\frac{\partial x_i}{\partial z_i} = 1 - \tanh^2 z_i \tag{6}$$

Derivative of E with respect to the sentiment classification matrix W_s

$$\frac{\partial E}{\partial W_s} = \sum_k \frac{\partial E}{\partial y_k} \frac{\partial y_k}{\partial z^s} \frac{\partial z^s}{\partial W_s} \tag{7}$$

Derivative of cost function:

$$\frac{\partial E}{\partial y} = \frac{t}{y} \tag{8}$$

Derivative of *softmax* function:

$$\frac{\partial y_k}{\partial z_i^s} = y_i(\delta_{ik} - y_k) \tag{9}$$

Derivative of input:

$$\frac{\partial z^s}{\partial W_s} = x \tag{10}$$

Combined:

$$\begin{aligned} \frac{\partial E}{\partial W_s} &= \sum_k \frac{t_k}{y_k} y_k(\delta_{ik} - y_i) x_j \\ &= x_j \sum_k t_k(\delta_{ik} - y_i) \\ &= x_j(y_i - t_i) \end{aligned} \tag{11}$$

Derivative of E with respect to the weight matrix W

For one training sentence:

$$\frac{\partial E}{\partial W} = \sum_k \frac{\partial E}{\partial y_k} \frac{\partial y_k}{\partial z_s} \frac{\partial z_s}{\partial x} \frac{\partial x}{\partial z} \frac{\partial z}{\partial W} \tag{12}$$

Derivative of input to $node_n$ w.r.t. activation of $node_{n-1}$:

$$\frac{\partial z}{\partial x} = W \tag{13}$$

Derivative of a node's activation w.r.t. its input:

$$\begin{aligned} \frac{\partial x}{\partial z} &= 1 - \tanh^2 z \\ f'(x) &= 1 - x^2 \\ f'\left(\begin{bmatrix} x^l \\ x^r \end{bmatrix}\right) &= 1 - \begin{bmatrix} x^l \\ x^r \end{bmatrix} \otimes \begin{bmatrix} x^l \\ x^r \end{bmatrix} \end{aligned} \tag{14}$$

Derivative of a node's input w.r.t. its weight matrix W:

$$\frac{\partial z}{\partial W} = x \tag{15}$$

Combined:

$$\begin{aligned} \delta^s &= {W_s}^T (y - t) \otimes f'(x_n) \\ \frac{\partial E}{\partial W} &= W^T \delta^s \otimes f'\left(\begin{bmatrix} x^l_{n-1} \\ x^r_{n-1} \end{bmatrix}\right) \begin{bmatrix} x^l_{n-1} \\ {x^r}_{n-1} \end{bmatrix}^T \end{aligned} \tag{16}$$

Derivative of E with respect to slice k of the tensor layer $V^{[k]}$

Top node ($node_n$):

$$\begin{aligned} \delta^s &= {W_s}^T (y - t) \otimes (1 - {x_n}^2) \\ \frac{\partial E_n}{\partial V^{[k]}} &= {\delta^s}_k \begin{bmatrix} {x^l}_{n-1} \\ {x^r}_{n-1} \end{bmatrix} \begin{bmatrix} x^l_{n-1} \\ {x^r}_{n-1} \end{bmatrix}^T \end{aligned} \tag{17}$$

Left child node ($node_{n-1}$):

$$\begin{aligned} \delta_n &= \delta^{s,n} \\ \delta_k^{n-1} &= \left(W^T \delta^n + S\right) \otimes f'\left(\left[\begin{array}{c} x_{n-1}^l \\ x_{n-1}^r \end{array}\right]\right) \\ S &= \sum_{k=1}^{d} \delta^n \left(V^{[k]} + \left(V^{[k]}\right)^T\right) \left[\begin{array}{c} x_{n-1}^l \\ x_{n-1}^r \end{array}\right] \\ \delta_l^{n-1} &= \delta_l^{s,n-1} + \delta^{n-1}[1:d] \\ \frac{\partial E_{n-1}}{\partial V^{[k]}} &= \frac{\partial E_n}{\partial V^{[k]}} + \delta_l^{n-1} \left[\begin{array}{c} x_{n-2}^l \\ {x^r}_{n-2} \end{array}\right]^T \end{aligned} \tag{18}$$