T-LSTM Forward and Back Propagation

May 25, 2015

1 Assumption

1. Parse tree is binary with ordered nodes
2. Only leaf nodes have words
3. Cross-entropy loss function

2 Forward Propagation

2.1 Non-Leaf Nodes

$$\hat{y} = \text{softmax}(W^{(s)}h + b^{(s)}) \tag{1}$$

$$h = o \odot \tanh(c) \tag{2}$$

$$c = i \odot u + f_l \odot c_l + f_r \odot c_r \tag{3}$$

$$f_l = a_l(U^{(l)}h_l + V^{(l)}h_r + b^{(f)}) \tag{4}$$

$$f_r = a_r(U^{(r)}h_l + V^{(r)}h_r + b^{(f)}) \tag{5}$$

$$u = a_u(U^{(u)}h_l + V^{(u)}h_r + b^{(u)}) \tag{6}$$

$$o = a_o(U^{(o)}h_l + V^{(o)}h_r + b^{(o)}) \tag{7}$$

$$i = a_i(U^{(i)}h_l + V^{(i)}h_r + b^{(i)}) \tag{8}$$

where a_j's are the activation functions.

2.2 Leaf Nodes

$$\hat{y} = \text{softmax}(W^{(s)}h + b^{(s)}) \tag{9}$$

$$h = o \odot \tanh(c) \tag{10}$$

$$c = i \odot u \tag{11}$$

$$u = a_u(W^{(u)}x + b^{(u)}) \tag{12}$$

$$o = a_o(W^{(o)}x + b^{(o)}) \tag{13}$$

$$i = a_i(W^{(i)}x + b^{(i)}) \tag{14}$$

3 Back Propagation

3.1 Error Flows

There are a total of 6 error outlets from each parent to each of its children. $h_{\text{child}} \to o$, $h_{\text{child}} \to i$, $h_{\text{child}} \to u$, $h_{\text{child}} \to f_l$, $h_{\text{child}} \to f_r$, $c_{\text{child}} \to c$.

Total Error at h : let the total error at h be denoted by e_h.

$$e_h = \frac{\partial J}{\partial h} + \delta_o U^{(o)} + \delta_i U^{(i)} + \delta_u U^{(u)} + \delta_l U^{(l)} + \delta_r U^{(r)} \tag{15}$$

where δ_j's are the input errors from parent node.
Note: In the above equation, it is assumed that the node under consideration is a left child of its parent. If the node is a right child, replace all U-parameters in the equation by the corresponding V-parameters.

Total Error at c : let the total error at c be denoted by e_c.

$$e_c = \frac{\partial J}{\partial c} + \delta_c \text{diag}(f_l) + e_h \frac{\partial h}{\partial c} \tag{16}$$

Note: In the above equation, it is assumed that the node under consideration is a left child of its parent. If the node is a right child, replace f_l by f_r.

Output Errors

let the output errors (going from node to its children be denoted by Δ_j's.

$$\Delta_o = e_h \operatorname{diag}(\tanh(c))\, \Sigma^{(o)} \tag{17}$$

$$\Delta_i = e_c \operatorname{diag}(u)\, \Sigma^{(i)} \tag{18}$$

$$\Delta_u = e_c \operatorname{diag}(i)\, \Sigma^{(u)} \tag{19}$$

$$\Delta_l = e_c \operatorname{diag}(c_l)\, \Sigma^{(l)} \tag{20}$$

$$\Delta_r = e_c \operatorname{diag}(c_r)\, \Sigma^{(r)} \tag{21}$$

$$\Delta_c = e_c \tag{22}$$

where $\Sigma^{(j)} = \operatorname{diag}(a'_j)$ and a'_j denotes the elementwise derivative of the activation function for a_j.

Derivatives wrt h and c

$$\frac{\partial J}{\partial h} = \frac{\partial J}{\partial \theta} W^{(s)} = (\hat{y} - y)^T W^{(s)} \tag{23}$$

where $\theta = W^{(s)}h + b^{(s)}$.

$$\frac{\partial J}{\partial c} = \frac{\partial J}{\partial h} \operatorname{diag}(o)\, \Sigma^{(c)} \tag{24}$$

where $\Sigma^{(c)} = d(\tanh(c))/dc$.

3.2 Parameter Derivatives

\# = defined only for non-leaf nodes
† = defined only for leaf nodes

Softmax Parameters

$$\frac{\partial J}{\partial b^{(s)}} = \frac{\partial J}{\partial \theta};\ \ \frac{\partial J}{\partial W^{(s)}} = h \frac{\partial J}{\partial \theta} \tag{25}$$

Bias Terms

$$\frac{\partial J}{\partial b^{(j)}} = \Delta_j;\ \ \frac{\partial J^{\#}}{\partial b^{(f)}} = \Delta_l + \Delta_r \tag{26}$$

where $j \in \{o, i, u\}$

U, V Parameters (#)

$$\frac{\partial J}{\partial U^{(j)}} = h_l \Delta_j;\ \ \frac{\partial J}{\partial V^{(j)}} = h_r \Delta_j \tag{27}$$

where $j \in \{o, i, u, l, r\}$

W Parameters (†)

$$\frac{\partial J}{\partial W^{(j)}} = x\Delta_j \tag{28}$$

where $j \in \{o, i, u\}$