Deep Recurrent Neural Network

Alex Adamson
aadamson@stanford.edu

May 13, 2015

1 Network specification

We borrow our specification from Irsoy and Cardie, 2014.

For $i > 1$, we have

$$\overrightarrow{h}_t^{(i)} = f(\underrightarrow{\overrightarrow{W}}^{(i)} \overrightarrow{h}_t^{(i-1)} + \underleftarrow{\overrightarrow{W}}^{(i)} \overleftarrow{h}_t^{(i-1)} + \overrightarrow{V}^{(i)} \overrightarrow{h}_{t-1}^{(i)} + \overrightarrow{b}^{(i)}) \tag{1}$$

$$\overleftarrow{h}_t^{(i)} = f(\underrightarrow{\overleftarrow{W}}^{(i)} \overrightarrow{h}_t^{(i-1)} + \underleftarrow{\overleftarrow{W}}^{(i)} \overleftarrow{h}_t^{(i-1)} + \overleftarrow{V}^{(i)} \overleftarrow{h}_{t=1}^{(i)} + \overleftarrow{b}^{(i)}) \tag{2}$$

and for $i = 1$ we have

$$\overrightarrow{h}_t^{(1)} = f(\overrightarrow{W}^{(1)} x_t + \overrightarrow{V}^{(1)} \overrightarrow{h}_{t-1}^{(1)} + \overrightarrow{b}^{(1)}) \tag{3}$$

$$\overleftarrow{h}_t^{(1)} = f(\overleftarrow{W}^{(1)} x_t + \overleftarrow{V}^{(1)} \overleftarrow{h}_{t+1}^{(1)} + \overleftarrow{b}^{(1)}) \tag{4}$$

We only connect the last layer (which we denote layer L) to the output layer:

$$y_t = g(\underrightarrow{U} \overrightarrow{h}_t^{(L)} + \underleftarrow{U} \overleftarrow{h}_t^{(L)} + c) \tag{5}$$

2 Backpropagation derivation

For convenience, we use error vector notation when deriving the backpropagation updates.

For our loss function, we choose categorical cross-entropy.

Let δ_t^y be the error vector propagated by the softmax unit at timestep t.

Let $\overrightarrow{\delta}_t^{(i)}$ be the error vector propagated by the forward hidden unit in layer i at timestep t.

Let $\overleftarrow{\delta}_t^{(i)}$ be the error vector propagated by the backward hidden unit in layer i at timestep t.

Let f^* be the function such that $f^*(f(x)) = f'(x)$ where f is as above.

Then:

$$\delta_t^y = \hat{y}_t - y_t \tag{6}$$

$$\overrightarrow{\delta}_t^{(i)} = f^*(\overrightarrow{h}_t^{(i)}) \circ ((\underrightarrow{\overrightarrow{W}}^{(i+1)})^T \overrightarrow{\delta}_t^{(i+1)} + (\underrightarrow{\overleftarrow{W}}^{(i+1)})^T \overleftarrow{\delta}_t^{(i+1)} + (\overrightarrow{V}^{(i)})^T \overrightarrow{\delta}_{t+1}^{(i)}) \tag{7}$$

$$\overleftarrow{\delta}_t^{(i)} = f^*(\overleftarrow{h}_t^{(i)}) \circ ((\underleftarrow{\overrightarrow{W}}^{(i+1)})^T \overrightarrow{\delta}_t^{(i+1)} + (\underleftarrow{\overleftarrow{W}}^{(i+1)})^T \overleftarrow{\delta}_t^{(i+1)} + (\overleftarrow{V}^{(i)})^T \overleftarrow{\delta}_{t-1}^{(i)}) \tag{8}$$

With these in hand, we can find the actual updates:

$$
\begin{aligned}
\frac{\partial J}{\partial \overrightarrow{U}} &= \delta^y (\overrightarrow{h}^{(L)})^T && (9)\\
\frac{\partial J}{\partial c} &= \delta^y \cdot \mathbf{1} && (10)\\
\frac{\partial J}{\partial \underrightarrow{\overrightarrow{W}}^{(i)}} &= \overrightarrow{\delta}^{(i)} (\overrightarrow{h}^{(i-1)})^T && (11)\\
\frac{\partial J}{\partial \underleftarrow{\overrightarrow{W}}^{(i)}} &= \overrightarrow{\delta}^{(i)} (\overleftarrow{h}^{(i-1)})^T && (12)\\
\frac{\partial J}{\partial \overrightarrow{b}^{(i)}} &= \overrightarrow{\delta}^{(i)} \cdot \mathbf{1} && (13)\\
\frac{\partial J}{\partial \underrightarrow{\overleftarrow{W}}^{(i)}} &= \overleftarrow{\delta}^{(i)} (\overrightarrow{h}^{(i-1)})^T && (14)\\
\frac{\partial J}{\partial \underleftarrow{\overleftarrow{W}}^{(i)}} &= \overleftarrow{\delta}^{(i)} (\overleftarrow{h}^{(i-1)})^T && (15)\\
\frac{\partial J}{\partial \overleftarrow{b}^{(i)}} &= \overleftarrow{\delta}^{(i)} \cdot \mathbf{1} && (16)\\
\frac{\partial J}{\partial \overrightarrow{V}^{(i)}} &= \sum_{t=1}^{T} \overrightarrow{\delta}_t^{(i)} (\overrightarrow{h}_{t-1}^{(i)})^T && (17)\\
\frac{\partial J}{\partial \overleftarrow{V}^{(i)}} &= \sum_{t=1}^{T} \overleftarrow{\delta}_t^{(i)} (\overleftarrow{h}_{t+1}^{(i)})^T && (18)\\
\frac{\partial J}{\partial \overrightarrow{W}^{(1)}} &= \overrightarrow{\delta}^{(1)} x_t^T && (19)\\
\frac{\partial J}{\partial \overleftarrow{W}^{(1)}} &= \overleftarrow{\delta}^{(1)} x_t^T && (20)
\end{aligned}
$$