Deep Recurrent Neural Network with Gated Recurrent Units

Alex Adamson
aadamson@stanford.edu

May 13, 2015

1 Network specification

This network is an extension of the network introduced in Irsoy and Cardie, 2014. The transfer functions between the units in a single layer now use gated recurrent methods.

For $i > 1$, we have

$$\begin{aligned}
\overrightarrow{z}_t^{(i)} &= f_2(\underrightarrow{\overrightarrow{Wz}}^{(i)}\overrightarrow{h}_t^{(i-1)} + \underleftarrow{\overrightarrow{Wz}}^{(i)}\overleftarrow{h}_t^{(i-1)} + \overrightarrow{Vz}^{(i)}\overrightarrow{h}_{t-1}^{(i)}) && (1)\\
\overrightarrow{r}_t^{(i)} &= f_2(\underrightarrow{\overrightarrow{Wr}}^{(i)}\overrightarrow{h}_t^{(i-1)} + \underleftarrow{\overrightarrow{Wr}}^{(i)}\overleftarrow{h}_t^{(i-1)} + \overrightarrow{Vr}^{(i)}\overrightarrow{h}_{t-1}^{(i)}) && (2)\\
\widetilde{\overrightarrow{h}}_t^{(i)} &= f(\underrightarrow{\overrightarrow{W}}^{(i)}\overrightarrow{h}_t^{(i-1)} + \underleftarrow{\overrightarrow{W}}^{(i)}\overleftarrow{h}_t^{(i-1)} + \overrightarrow{r}_t^{(i)} \circ \overrightarrow{V}^{(i)}\overrightarrow{h}_{t-1}^{(i)}) && (3)\\
\overrightarrow{h}_t^{(i)} &= \overrightarrow{z}_t^{(i)} \circ \overrightarrow{h}_{t-1}^{(i)} + (1 - \overrightarrow{z}_t^{(i)}) \circ \widetilde{\overrightarrow{h}}_t^{(i)} && (4)\\
\overleftarrow{r}_t^{(i)} &= f_2(\underrightarrow{\overleftarrow{Wr}}^{(i)}\overrightarrow{h}_t^{(i-1)} + \underleftarrow{\overleftarrow{Wr}}^{(i)}\overleftarrow{h}_t^{(i-1)} + \overleftarrow{Vr}^{(i)}\overleftarrow{h}_{t+1}^{(i)}) && (5)\\
\overleftarrow{z}_t^{(i)} &= f_2(\underrightarrow{\overleftarrow{Wz}}^{(i)}\overrightarrow{h}_t^{(i-1)} + \underleftarrow{\overleftarrow{Wz}}^{(i)}\overleftarrow{h}_t^{(i-1)} + \overleftarrow{Vz}^{(i)}\overleftarrow{h}_{t+1}^{(i)}) && (6)\\
\widetilde{\overleftarrow{h}}_t^{(i)} &= f(\underrightarrow{\overleftarrow{W}}^{(i)}\overrightarrow{h}_t^{(i-1)} + \underleftarrow{\overleftarrow{W}}^{(i)}\overleftarrow{h}_t^{(i-1)} + \overleftarrow{r}_t^{(i)} \circ \overleftarrow{V}^{(i)}\overleftarrow{h}_{t=1}^{(i)}) && (7)\\
\overleftarrow{h}_t^{(i)} &= \overleftarrow{z}_t^{(i)} \circ \overleftarrow{h}_{t+1}^{(i)} + (1 - \overleftarrow{z}_t^{(i)}) \circ \widetilde{\overleftarrow{h}}_t^{(i)} && (8)
\end{aligned}$$

and for $i = 1$ we have

$$\begin{aligned}
\overrightarrow{h}_t^{(1)} &= f(\overrightarrow{W}^{(1)}x_t + \overrightarrow{V}^{(1)}\overrightarrow{h}_{t-1}^{(1)} + \overrightarrow{b}^{(1)}) && (9)\\
\overleftarrow{h}_t^{(1)} &= f(\overleftarrow{W}^{(1)}x_t + \overleftarrow{V}^{(1)}\overleftarrow{h}_{t+1}^{(1)} + \overleftarrow{b}^{(1)}) && (10)
\end{aligned}$$

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

$$y_t = g(\underrightarrow{U}\overrightarrow{h}_t^{(L)} + \underleftarrow{U}\overleftarrow{h}_t^{(L)} + c) \tag{11}$$

2 Backpropagation derivation

We again use error vector notation when deriving the gradients via backpropagation.

Let δ_t^y be the error vector propagated by the softmax unit at timestep t. Note that this is unchanged from the original network.

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.

We first derive $\overrightarrow{\delta}_t^{(i)}$. Let f_2^* be a function such that $f_2^*(f_2(x)) = f_2'(x)$. Let f^* be the function such that $f^*(f(x)) = f'(x)$. Then,

$$\overrightarrow{\delta}_t^{(i)} = \frac{\partial \overrightarrow{h}_t^{(i+1)}}{\partial \overrightarrow{h}_t^{(i)}} \frac{\partial J}{\partial \overrightarrow{h}_t^{(i+1)}} + \frac{\partial \overleftarrow{h}_t^{(i+1)}}{\partial \overrightarrow{h}_t^{(i)}} \frac{\partial J}{\partial \overleftarrow{h}_t^{(i+1)}} + \frac{\partial \overrightarrow{h}_{t+1}^{i}}{\partial \overrightarrow{h}_t^{(i)}} \frac{\partial J}{\partial \overrightarrow{h}_{t+1}^{i}}$$

$$\begin{aligned}
\frac{\partial \overrightarrow{h}_t^{(i+1)}}{\partial \overrightarrow{h}_t^{(i)}} \frac{\partial J}{\partial \overrightarrow{h}_t^{(i+1)}} &= \frac{\partial \overrightarrow{h}_t^{(i+1)}}{\partial \overrightarrow{h}_t^{(i)}} \cdot \overrightarrow{\delta}_t^{(i+1)} \\
&= \left(\frac{\partial \overrightarrow{z}_t^{(i+1)}}{\partial \overrightarrow{h}_t^{(i)}} \circ \overrightarrow{h}_{t-1}^{(i+1)} + \frac{\partial (1 - \overrightarrow{z}_t^{(i+1)})}{\partial \overrightarrow{h}_t^{(i)}} \circ \widetilde{\overrightarrow{h}}_t^{(i+1)} + (1 - \overrightarrow{z}_t^{(i+1)}) \circ \frac{\partial \widetilde{\overrightarrow{h}}_t^{(i+1)}}{\partial \overrightarrow{h}_t^{(i)}}\right) \cdot \overrightarrow{\delta}_t^{(i+1)} \\
&= \left(\frac{\partial \overrightarrow{z}_t^{(i+1)}}{\partial \overrightarrow{h}_t^{(i)}} \circ (\overrightarrow{h}_{t-1}^{(i+1)} - \widetilde{\overrightarrow{h}}_t^{(i+1)}) + (1 - \overrightarrow{z}_t^{(i+1)}) \circ \frac{\partial \widetilde{\overrightarrow{h}}_t^{(i+1)}}{\partial \overrightarrow{h}_t^{(i)}}\right) \cdot \overrightarrow{\delta}_t^{(i+1)} \\
&= \left((f_2^*(\overrightarrow{z}_t^{(i+1)}) \circ (\overrightarrow{W_z}^{(i+1)})^T) \circ (\overrightarrow{h}_{t-1}^{(i+1)} - \widetilde{\overrightarrow{h}}_t^{(i+1)}) + (1 - \overrightarrow{z}_t^{(i+1)}) \circ \frac{\partial \widetilde{\overrightarrow{h}}_t^{(i+1)}}{\partial \overrightarrow{h}_t^{(i)}}\right) \cdot \overrightarrow{\delta}_t^{(i+1)} \\
&= \left((f_2^*(\overrightarrow{z}_t^{(i+1)}) \circ (\overrightarrow{W_z}^{(i+1)})^T) \circ (\overrightarrow{h}_{t-1}^{(i+1)} - \widetilde{\overrightarrow{h}}_t^{(i+1)}) + (1 - \overrightarrow{z}_t^{(i+1)}) \circ f^*(\widetilde{\overrightarrow{h}}_t^{(i+1)}) \circ (\overrightarrow{W_r}^{(i+1)})^T\right) \cdot \overrightarrow{\delta}_t^{(i+1}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \overleftarrow{h}_t^{(i+1)}}{\partial \overrightarrow{h}_t^{(i)}} \frac{\partial J}{\partial \overleftarrow{h}_t^{(i+1)}} &= \frac{\partial \overleftarrow{h}_t^{(i+1)}}{\partial \overrightarrow{h}_t^{(i)}} \cdot \overleftarrow{\delta}_t^{(i+1)} \\
&= \left(\frac{\partial \overleftarrow{z}_t^{(i+1)}}{\partial \overrightarrow{h}_t^{(i)}} \circ \overleftarrow{h}_{t+1}^{(i+1)} + \frac{\partial (1 - \overleftarrow{z}_t^{(i+1)})}{\partial \overrightarrow{h}_t^{(i)}} \circ \widetilde{\overleftarrow{h}}_t^{(i+1)} + (1 - \overleftarrow{z}_t^{(i+1)}) \circ \frac{\partial \widetilde{\overleftarrow{h}}_t^{(i+1)}}{\partial \overrightarrow{h}_t^{(i)}}\right) \cdot \overleftarrow{\delta}_t^{(i+1)} \\
&= \left(\frac{\partial \overleftarrow{z}_t^{(i+1)}}{\partial \overrightarrow{h}_t^{(i)}} \circ (\overleftarrow{h}_{t+1}^{(i+1)} - \widetilde{\overleftarrow{h}}_t^{(i+1)}) + (1 - \overleftarrow{z}_t^{(i+1)}) \circ \frac{\partial \widetilde{\overleftarrow{h}}_t^{(i+1)}}{\partial \overrightarrow{h}_t^{(i)}}\right) \cdot \overleftarrow{\delta}_t^{(i+1)} \\
&= \left((f_2^*(\overleftarrow{z}_t^{(i+1)}) \circ (\overleftarrow{W_z}^{(i+1)})^T) \circ (\overleftarrow{h}_{t+1}^{(i+1)} - \widetilde{\overleftarrow{h}}_t^{(i+1)}) + (1 - \overleftarrow{z}_t^{(i+1)}) \circ \frac{\partial \widetilde{\overleftarrow{h}}_t^{(i+1)}}{\partial \overrightarrow{h}_t^{(i)}}\right) \cdot \overleftarrow{\delta}_t^{(i+1)} \\
&= \left((f_2^*(\overleftarrow{z}_t^{(i+1)}) \circ (\overleftarrow{W_z}^{(i+1)})^T) \circ (\overleftarrow{h}_{t+1}^{(i+1)} - \widetilde{\overleftarrow{h}}_t^{(i+1)}) + (1 - \overleftarrow{z}_t^{(i+1)}) \circ f^*(\widetilde{\overleftarrow{h}}_t^{(i+1)}) \circ (\overleftarrow{W_r}^{(i+1)})^T\right) \cdot \overleftarrow{\delta}_t^{(i+1}
\end{aligned}$$

$$
\begin{aligned}
\frac{\partial \overrightarrow{h}_{t+1}^{(i)}}{\partial \overrightarrow{h}_{t}^{(i)}} \frac{\partial J}{\partial \overrightarrow{h}_{t+1}^{(i)}} &= \frac{\partial \overrightarrow{h}_{t+1}^{(i0}}{\partial \overrightarrow{h}_{t}^{(i)}} \cdot \overrightarrow{\delta}_{t+1}^{(i)} \\
&= (\overrightarrow{z}_{t+1}^{(i)} + \frac{\partial \overrightarrow{z}_{t+1}^{(i)}}{\partial \overrightarrow{h}_{t}^{(i)}} \circ \overrightarrow{h}_{t}^{(i)} + \frac{\partial (1 - \overrightarrow{z}_{t+1}^{(i)})}{\partial \overrightarrow{h}_{t}^{(i)}} \circ \widetilde{\overrightarrow{h}}_{t+1}^{(i)} + (1 - \overrightarrow{z}_{t+1}^{(i)}) \circ \frac{\partial \widetilde{\overrightarrow{h}}_{t+1}^{(i)}}{\partial \overrightarrow{h}_{t}^{(i)}}) \cdot \overrightarrow{\delta}_{t+1}^{(i)} \\
&= (\overrightarrow{z}_{t+1}^{(i)} + \frac{\partial \overrightarrow{z}_{t+1}^{(i)}}{\partial \overrightarrow{h}_{t}^{(i)}} \circ (\overrightarrow{h}_{t}^{(i)} - \widetilde{\overrightarrow{h}}_{t+1}^{(i)}) + (1 - \overrightarrow{z}_{t+1}^{(i)}) \circ \frac{\partial \widetilde{\overrightarrow{h}}_{t+1}^{(i)}}{\partial \overrightarrow{h}_{t}^{(i)}}) \cdot \overrightarrow{\delta}_{t+1}^{(i)} \\
&= (\overrightarrow{z}_{t+1}^{(i)} + (f_2^*(\overrightarrow{z}_{t+1}^{(i)}) \circ (\overrightarrow{Vz}^{(i)})^T) \circ (\overrightarrow{h}_{t}^{(i)} - \widetilde{\overrightarrow{h}}_{t+1}^{(i)}) + (1 - \overrightarrow{z}_{t+1}^{(i)}) \circ \frac{\partial \widetilde{\overrightarrow{h}}_{t+1}^{(i)}}{\partial \overrightarrow{h}_{t}^{(i)}}) \cdot \overrightarrow{\delta}_{t+1}^{(i)} \\
&= (\overrightarrow{z}_{t+1}^{(i)} + (f_2^*(\overrightarrow{z}_{t+1}^{(i)}) \circ (\overrightarrow{Vz}^{(i)})^T) \circ (\overrightarrow{h}_{t}^{(i)} - \widetilde{\overrightarrow{h}}_{t+1}^{(i)}) \\
&\quad + (1 - \overrightarrow{z}_{t+1}^{(i)}) \circ f^*(\widetilde{\overrightarrow{h}}_{t+1}^{(i)}) \circ \frac{\partial (\overrightarrow{r}_{t+1}^{(i)} \circ \overrightarrow{V}^{(i)} \overrightarrow{h}_{t}^{(i)})}{\partial \overrightarrow{h}_{t}^{(i)}}) \cdot \overrightarrow{\delta}_{t+1}^{(i)} \\
&= (\overrightarrow{z}_{t+1}^{(i)} + (f_2^*(\overrightarrow{z}_{t+1}^{(i)}) \circ (\overrightarrow{Vz}^{(i)})^T) \circ (\overrightarrow{h}_{t}^{(i)} - \widetilde{\overrightarrow{h}}_{t+1}^{(i)}) \\
&\quad + (1 - \overrightarrow{z}_{t+1}^{(i)}) \circ f^*(\widetilde{\overrightarrow{h}}_{t+1}^{(i)}) \circ (\frac{\partial \overrightarrow{r}_{t+1}^{(i)}}{\partial \overrightarrow{h}_{t}^{(i)}} + \frac{\partial \overrightarrow{V}^{(i)} \overrightarrow{h}_{t}^{(i)}}{\partial \overrightarrow{h}_{t}^{(i)}})) \cdot \overrightarrow{\delta}_{t+1}^{(i)} \\
&= (\overrightarrow{z}_{t+1}^{(i)} + (f_2^*(\overrightarrow{z}_{t+1}^{(i)}) \circ (\overrightarrow{Vz}^{(i)})^T) \circ (\overrightarrow{h}_{t}^{(i)} - \widetilde{\overrightarrow{h}}_{t+1}^{(i)}) \\
&\quad + (1 - \overrightarrow{z}_{t+1}^{(i)}) \circ f^*(\widetilde{\overrightarrow{h}}_{t+1}^{(i)}) \circ (f_2^*(\overrightarrow{r}_{t+1}^{(i)}) \circ (\overrightarrow{Vr}^{(i)})^T + (\overrightarrow{V}^{(i)})^T)) \cdot \overrightarrow{\delta}_{t+1}^{(i)}
\end{aligned}
$$