LSTM Derivations

Minh-Thang Luong
lmthang@stanford.edu

Abigail See
abisee@stanford.edu

1 LSTM Architecture

An LSTM block at layer $l \in \{1, \dots L\}$ and time $t \in \{1, \dots T\}$ consists of:

- The hidden state $\boldsymbol{h_t^l} \in \mathbb{R}^n$
- The memory cell $\boldsymbol{c_t^l} \in \mathbb{R}^n$
- The input gate $\boldsymbol{i_t^l} \in [0,1]^n$
- The forget gate $\boldsymbol{f_t^l} \in [0,1]^n$
- The output gate $\boldsymbol{o_t^l} \in [0,1]^n$
- The input modulation gate $\boldsymbol{\hat{h}_t^l} \in [0,1]^n$

We call n the LSTM block size.

2 Forward Propagation

We use the formulation of [Zaremba et al., 2014]. For a single LSTM block at layer l and time t, the new hidden state $\boldsymbol{h_t^l}$ and memory cell $\boldsymbol{c_t^l}$ are calculated from $\boldsymbol{h_t^{l-1}}$, $\boldsymbol{h_{t-1}^l}$ and $\boldsymbol{c_{t-1}^l}$ like so:

$$\begin{pmatrix} \boldsymbol{i_t^l} \\ \boldsymbol{f_t^l} \\ \boldsymbol{o_t^l} \\ \boldsymbol{\hat{h}_t^l} \end{pmatrix} = \begin{pmatrix} \mathrm{sigm} \\ \mathrm{sigm} \\ \mathrm{sigm} \\ \tanh \end{pmatrix} \boldsymbol{T}_{4n \times 2n} \begin{bmatrix} \boldsymbol{h_t^{l-1}} \\ \boldsymbol{h_{t-1}^l} \end{bmatrix} \tag{1}$$

$$\boldsymbol{c_t^l} = \boldsymbol{f_t^l} \circ \boldsymbol{c_{t-1}^l} + \boldsymbol{i_t^l} \circ \boldsymbol{\hat{h}_t^l} \tag{2}$$

$$\boldsymbol{h_t^l} = \boldsymbol{o_t^l} \circ \tanh(\boldsymbol{c_t^l}) \tag{3}$$

where sigm and tanh are applied element-wise, $\circ$ denotes element-wise multiplication, and $\boldsymbol{T}_{4n \times 2n}$ is a $4n \times 2n$ matrix of weights that depends on l but not t.[1] If $l = 1$ then $\boldsymbol{h_t^{l-1}}$ is the input vector x_t. If $t = 1$ then $\boldsymbol{h_{t-1}^l}$ and $\boldsymbol{c_{t-1}^l}$ are taken to be zero.

[1]Note: Sometimes these equations are written omitting the superscript l and writing $\boldsymbol{h_t^{l-1}}$ as x_t, but for the purposes of deriving the back-propagation equations, we need to refer to the layer l explicitly.

3 Backward Propagation

3.1 Definitions

In this section we define some additional notation that will help us to derive the necessary back-propagation equations.

Definition 1. *Let U and V refer to the $n \times n$ weight matrices corresponding to the following portions of $\boldsymbol{T}_{4n\times 2n}$:*

$$\boldsymbol{T}_{4n\times 2n} = \begin{bmatrix} U_i & V_i \\ U_f & V_f \\ U_o & V_o \\ U_{\hat{h}} & V_{\hat{h}} \end{bmatrix} \tag{4}$$

In particular, we will use the superscript l to denote these matrices used to calculate layer l in Equation (1).

Definition 2. *For all $l \in \{1, \ldots, L\}$ and $t \in \{1, \ldots, T\}$, define the* weighted inputs

$$z_i^l(t) = U_i \boldsymbol{h_t^{l-1}} + V_i \boldsymbol{h_{t-1}^l} \qquad z_f^l(t) = U_f \boldsymbol{h_t^{l-1}} + V_f \boldsymbol{h_{t-1}^l}$$
$$z_o^l(t) = U_o \boldsymbol{h_t^{l-1}} + V_o \boldsymbol{h_{t-1}^l} \qquad z_{\hat{h}}^l(t) = U_{\hat{h}} \boldsymbol{h_t^{l-1}} + V_{\hat{h}} \boldsymbol{h_{t-1}^l}$$

so that

$$\begin{pmatrix} \boldsymbol{i_t^l} \\ \boldsymbol{f_t^l} \\ \boldsymbol{o_t^l} \\ \boldsymbol{\hat{h}_t^l} \end{pmatrix} = \begin{pmatrix} \text{sigm} \\ \text{sigm} \\ \text{sigm} \\ \tanh \end{pmatrix} \begin{pmatrix} z_i^l(t) \\ z_f^l(t) \\ z_o^l(t) \\ z_{\hat{h}}^l(t) \end{pmatrix} \tag{5}$$

where sigm *and* tanh *are applied element-wise. We call $z_i^l(t)$ the* weighted input *to the input gate $\boldsymbol{i_t^l}$.*

Definition 3. *Define the* error *of the input, forget, output and input modulation gates at layer l and time t to be*

$$\boldsymbol{\delta_i^l(t)} = \frac{\partial L}{\partial z_i^l(t)} \qquad \boldsymbol{\delta_f^l(t)} = \frac{\partial L}{\partial z_f^l(t)}$$
$$\boldsymbol{\delta_o^l(t)} = \frac{\partial L}{\partial z_o^l(t)} \qquad \boldsymbol{\delta_{\hat{h}}^l(t)} = \frac{\partial L}{\partial z_{\hat{h}}^l(t)}$$

where L is the loss function. Note: $\boldsymbol{\delta_i^l(t)}$ is the partial derivative of L with respect to the weighted input $z_i^l(t)$, not *$\boldsymbol{i_t^l}$.*

Definition 4. *Define the* error *of the hidden state and cell and layer l and time t to be*

$$\boldsymbol{\delta_h^l(t)} = \frac{\partial L}{\partial \boldsymbol{h_t^l}} \qquad \boldsymbol{\delta_c^l(t)} = \frac{\partial L}{\partial \boldsymbol{c_t^l}} \tag{6}$$

where L is the loss function.

3.2 Derivations

In this section we will derive expressions for $\boldsymbol{\delta^l_h(t)}$, $\boldsymbol{\delta^l_c(t)}$, $\boldsymbol{\delta^l_i(t)}$, $\boldsymbol{\delta^l_f(t)}$, $\boldsymbol{\delta^l_o(t)}$, and $\boldsymbol{\delta^l_{\hat{h}}(t)}$ in terms of the $\boldsymbol{\delta}$ values for the $(l+1,t)$ and $(l,t+1)$ blocks. These expressions will enable us to do back-propagation through time and layers. If you are not interested in the derivations, skip ahead to Section 3.3 to see the final back-propagation equations.

Lemma 1. *For $l \in \{1,\ldots,L-1\}$ and $t \in \{1,\ldots,T-1\}$,*

$$\boldsymbol{\delta^l_h(t)} = \begin{bmatrix} U_i^\top & U_f^\top & U_o^\top & U_{\hat{h}}^\top & V_i^\top & V_f^\top & V_o^\top & V_{\hat{h}}^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{\delta^l_i(t+1)} \\ \boldsymbol{\delta^l_f(t+1)} \\ \boldsymbol{\delta^l_o(t+1)} \\ \boldsymbol{\delta^l_{\hat{h}}(t+1)} \\ \boldsymbol{\delta^{l+1}_i(t)} \\ \boldsymbol{\delta^{l+1}_f(t)} \\ \boldsymbol{\delta^{l+1}_o(t)} \\ \boldsymbol{\delta^{l+1}_{\hat{h}}(t)} \end{bmatrix} \tag{7}$$

where each of the U and V matrices are with respect to layer l.

Note the left matrix in the multiplication has dimensions $n \times 8n$, the right matrix $8n \times n$, and $\boldsymbol{\delta^l_h(t)}$ is $n \times 1$.

Proof. We prove this element-wise. For any $j = 1, \ldots n$:

$$\boldsymbol{\delta^l_h(t)}_j = \frac{\partial L}{\partial (\boldsymbol{h^l_t})_j} \qquad \text{(definition of } \boldsymbol{\delta^l_h(t)}\text{)}$$

Now, because $\boldsymbol{h^l_t}$ affects $(z_i, z_f, z_o, z_{\hat{h}})$ for $(l, t+1)$ and $(l+1, t)$, we take the chain rule over these eight variables. Therefore the above equation can be written as

$$\begin{aligned}\sum_{k=1}^{n} \Bigg(& \frac{\partial L}{\partial z^l_i(t+1)_k} \frac{\partial z^l_i(t+1)_k}{\partial (\boldsymbol{h^l_t})_j} + \frac{\partial L}{\partial z^l_f(t+1)_k} \frac{\partial z^l_f(t+1)_k}{\partial (\boldsymbol{h^l_t})_j} \\ + & \frac{\partial L}{\partial z^l_o(t+1)_k} \frac{\partial z^l_o(t+1)_k}{\partial (\boldsymbol{h^l_t})_j} + \frac{\partial L}{\partial z^l_{\hat{h}}(t+1)_k} \frac{\partial z^l_{\hat{h}}(t+1)_k}{\partial (\boldsymbol{h^l_t})_j} \\ + & \frac{\partial L}{\partial z^{l+1}_i(t)_k} \frac{\partial z^{l+1}_i(t)_k}{\partial (\boldsymbol{h^l_t})_j} + \frac{\partial L}{\partial z^{l+1}_f(t)_k} \frac{\partial z^{l+1}_f(t)_k}{\partial (\boldsymbol{h^l_t})_j} \\ + & \frac{\partial L}{\partial z^{l+1}_o(t)_k} \frac{\partial z^{l+1}_o(t)_k}{\partial (\boldsymbol{h^l_t})_j} + \frac{\partial L}{\partial z^{l+1}_{\hat{h}}(t)_k} \frac{\partial z^{l+1}_{\hat{h}}(t)_k}{\partial (\boldsymbol{h^l_t})_j} \Bigg)\end{aligned}$$

First note that the first of each pair is some $\boldsymbol{\delta}$ e.g.

$$\frac{\partial L}{\partial z^l_i(t+1)_k} = \boldsymbol{\delta^l_i(t+1)} \qquad \text{(by definition)}$$

The second of each pair can be evaluated like so:

$$\begin{aligned}
\frac{\partial z_i^l(t+1)_k}{\partial(\boldsymbol{h_t^l})_j} &= \frac{\partial}{\partial(\boldsymbol{h_t^l})_j}\left(U_i^l \boldsymbol{h_{t+1}^l} + V_i^l \boldsymbol{h_t^l}\right)_k && \text{(definition of } z_i^l(t+1)\text{)} \\
&= \frac{\partial}{\partial(\boldsymbol{h_t^l})_j}\left(\sum_{m=1}^{n}(V_i^l)_{km}(\boldsymbol{h_t^l})_m\right) && (U_i^l \boldsymbol{h_{t+1}^l} \text{ does not depend on } \boldsymbol{h_t^l}) \\
&= (V_i^l)_{kj} && \text{(expression equals 0 except when } m=j\text{)}
\end{aligned}$$

so the first of the eight sums can be written as

$$\sum_{k=1}^{n}\frac{\partial L}{\partial z_i^l(t+1)_k}\frac{\partial z_i^l(t+1)_k}{\partial(\boldsymbol{h_t^l})_j} = \sum_{k=1}^{n}\boldsymbol{\delta_i^l(t+1)}(V_i^l)_{kj} = \left[(V_i^l)^\top \boldsymbol{\delta_i^l(t+1)}\right]_j \tag{8}$$

Finding similar expressions for the other seven sums, we obtain

$$\begin{aligned}
\boldsymbol{\delta_h^l(t)} \quad = \quad & (U_i^l)^\top \delta_i^l(t+1) \quad + \quad (U_f^l)^\top \delta_f^l(t+1) \quad + \quad (U_o^l)^\top \delta_o^l(t+1) \quad + \quad (U_{\hat{h}}^l)^\top \delta_{\hat{h}}^l(t+1) \\
+ \quad & (V_i^l)^\top \delta_i^{l+1}(t) \quad + \quad (V_f^l)^\top \delta_f^{l+1}(t) \quad + \quad (V_o^l)^\top \delta_o^{l+1}(t) \quad + \quad (V_{\hat{h}}^l)^\top \delta_{\hat{h}}^{l+1}(t)
\end{aligned}$$

□

Lemma 2. *For $l \in \{1,\dots,L\}$ and $t \in \{1,\dots,T-1\}$,*

$$\boldsymbol{\delta_c^l(t)} = \boldsymbol{\delta_c^l(t+1)} \circ f_{t+1}^l + \boldsymbol{\delta_h^l(t)} \circ \boldsymbol{o_t^l} \circ \tanh'(\boldsymbol{c_t^l}) \tag{9}$$

Proof. We prove this element-wise. For any $j = 1,\dots n$:

$$\begin{aligned}
\boldsymbol{\delta_c^l(t)}_j &= \frac{\partial L}{\partial(\boldsymbol{c_t^l})_j} && \text{(definition of } \boldsymbol{\delta_c^l(t)}\text{)} \\
&= \sum_{k=1}^{n}\frac{\partial L}{\partial(\boldsymbol{c_{t+1}^l})_k}\frac{\partial(\boldsymbol{c_{t+1}^l})_k}{\partial(\boldsymbol{c_t^l})_j} + \sum_{k=1}^{n}\frac{\partial L}{\partial(\boldsymbol{h_t^l})_k}\frac{\partial(\boldsymbol{h_t^l})_k}{\partial(\boldsymbol{c_t^l})_j} && \text{(chain rule)}
\end{aligned}$$

The second equality follows from the fact that $\boldsymbol{c_t^l}$ affects $\boldsymbol{h_t^l}$ and $\boldsymbol{c_{t+1}^l}$. For the first part of the expression, note that

$$\begin{aligned}
\frac{\partial(\boldsymbol{c_{t+1}^l})_k}{\partial(\boldsymbol{c_t^l})_j} &= \frac{\partial}{\partial(\boldsymbol{c_t^l})_j}\left(\boldsymbol{f_{t+1}^l} \circ \boldsymbol{c_t^l} + \boldsymbol{i_{t+1}^l} \circ \boldsymbol{\hat{h}_{t+1}^l}\right)_k && \text{(by definition of } \boldsymbol{c_{t+1}^l}\text{)} \\
&= \begin{cases} \boldsymbol{f_{t+1}^l} & \text{if } k=j \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}$$

For the second part, note that

$$\begin{aligned}
\frac{\partial(\boldsymbol{h_t^l})_k}{\partial(\boldsymbol{c_t^l})_j} &= \frac{\partial}{\partial(\boldsymbol{c_t^l})_j}\left(\boldsymbol{o_t^l} \circ \tanh(\boldsymbol{c_t^l})\right)_k && \text{(by definition of } \boldsymbol{h_t^l}\text{)} \\
&= \begin{cases} (\boldsymbol{o_t^l})_j \tanh'(\boldsymbol{c_t^l})_j & \text{if } k=j \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}$$

Combining the previous three equations and using the definitions of $\boldsymbol{\delta^l_c(t+1)}$ and $\boldsymbol{\delta^l_h(t)}$, we obtain

$$\boldsymbol{\delta^l_c(t)}_j = \boldsymbol{\delta^l_c(t+1)}_j(\boldsymbol{f^l_{t+1}})_j + \boldsymbol{\delta^l_h(t)}_j(\boldsymbol{o^l_t})_j \tanh'(\boldsymbol{c^l_t})_j \tag{10}$$

$\square$

Lemma 3. *For $l \in \{1, \dots, L\}$ and $t \in \{1, \dots, T\}$,*

$$\boldsymbol{\delta^l_i(t)} = \boldsymbol{\delta^l_c(t)} \circ \mathrm{sigm}'(z^l_i(t)) \circ \boldsymbol{\hat{h}^l_t} \tag{11}$$

Proof. We prove this element-wise. For any $j = 1, \dots n$:

$$\begin{aligned}
\boldsymbol{\delta^l_i(t)}_j &= \frac{\partial L}{\partial z^l_i(t)_j} && \text{(definition of } \boldsymbol{\delta^l_i(t)}\text{)}\\
&= \sum_{k=1}^{n} \frac{\partial L}{\partial(\boldsymbol{c^l_t})_k}\frac{\partial(\boldsymbol{c^l_t})_k}{\partial z^l_i(t)_j} && \text{(chain rule)}\\
&= \sum_{k=1}^{n} \boldsymbol{\delta^l_c(t)}_k \frac{\partial}{\partial z^l_i(t)_j}\left(\boldsymbol{f^l_t} \circ \boldsymbol{c^l_{t-1}} + \boldsymbol{i^l_t} \circ \boldsymbol{\hat{h}^l_t}\right)_k && \text{(definition of } \boldsymbol{\delta^l_c(t)} \text{ and } \boldsymbol{c^l_t}\text{)}\\
&= \boldsymbol{\delta^l_c(t)}_j \frac{\partial}{\partial z^l_i(t)_j}\left(\boldsymbol{i^l_t} \circ \boldsymbol{\hat{h}^l_t}\right)_j && \text{(expression equals 0 except when } k = j\text{)}\\
&= \boldsymbol{\delta^l_c(t)}_j \,\mathrm{sigm}'(z^l_i(t))_j(\boldsymbol{\hat{h}^l_t})_j && \text{(definition of } \boldsymbol{i^l_t} \text{ in terms of } z^l_i(t)\text{)}
\end{aligned}$$

Note that for the second equality we took the chain rule with respect to the elements of $\boldsymbol{c^l_t}$, because $\boldsymbol{i^l_t}$ affects $\boldsymbol{c^l_t}$. $\square$

Lemma 4. *For $l \in \{1, \dots, L\}$ and $t \in \{1, \dots, T\}$,*

$$\boldsymbol{\delta^l_f(t)} = \boldsymbol{\delta^l_c(t)} \circ \mathrm{sigm}'(z^l_f(t)) \circ \boldsymbol{c^l_{t-1}} \tag{12}$$

Proof. We prove this element-wise. For any $j = 1, \dots n$:

$$\begin{aligned}
\boldsymbol{\delta^l_f(t)}_j &= \frac{\partial L}{\partial z^l_f(t)_j} && \text{(definition of } \boldsymbol{\delta^l_f(t)}\text{)}\\
&= \sum_{k=1}^{n} \frac{\partial L}{\partial(\boldsymbol{c^l_t})_k}\frac{\partial(\boldsymbol{c^l_t})_k}{\partial z^l_f(t)_j} && \text{(chain rule)}\\
&= \sum_{k=1}^{n} \boldsymbol{\delta^l_c(t)}_k \frac{\partial}{\partial z^l_f(t)_j}\left(\boldsymbol{f^l_t} \circ \boldsymbol{c^l_{t-1}} + \boldsymbol{i^l_t} \circ \boldsymbol{\hat{h}^l_t}\right)_k && \text{(definition of } \boldsymbol{\delta^l_c(t)} \text{ and } \boldsymbol{c^l_t}\text{)}\\
&= \boldsymbol{\delta^l_c(t)}_j \frac{\partial}{\partial z^l_f(t)_j}\left(\boldsymbol{f^l_t} \circ \boldsymbol{c^l_{t-1}}\right)_j && \text{(expression equals 0 except when } k = j\text{)}\\
&= \boldsymbol{\delta^l_c(t)}_j \,\mathrm{sigm}'(z^l_f(t))_j(\boldsymbol{c^l_{t-1}})_j && \text{(definition of } \boldsymbol{f^l_t} \text{ in terms of } z^l_f(t)\text{)}
\end{aligned}$$

Note that for the second equality we took the chain rule with respect to the elements of $\boldsymbol{c^l_t}$, because $\boldsymbol{f^l_t}$ affects $\boldsymbol{c^l_t}$. $\square$

Lemma 5. *For $l \in \{1, \ldots, L\}$ and $t \in \{1, \ldots, T\}$,*

$$\boldsymbol{\delta_o^l(t)} = \boldsymbol{\delta_h^l(t)} \circ \operatorname{sigm}'(z_o^l(t)) \circ \tanh(\boldsymbol{c_t^l}) \tag{13}$$

Proof. We prove this element-wise. For any $j = 1, \ldots n$:

$$\begin{aligned}
\boldsymbol{\delta_o^l(t)}_j &= \frac{\partial L}{\partial z_o^l(t)_j} && \text{(definition of } \boldsymbol{\delta_o^l(t)}\text{)} \\
&= \sum_{k=1}^{n} \frac{\partial L}{\partial(\boldsymbol{h_t^l})_k} \frac{\partial(\boldsymbol{h_t^l})_k}{\partial z_o^l(t)_j} && \text{(chain rule)} \\
&= \sum_{k=1}^{n} \boldsymbol{\delta_h^l(t)}_k \frac{\partial}{\partial z_o^l(t)_j} \left(\boldsymbol{o_t^l} \circ \tanh(\boldsymbol{c_t^l})\right)_k && \text{(definition of } \boldsymbol{\delta_h^l(t)} \text{ and } \boldsymbol{h_t^l}\text{)} \\
&= \boldsymbol{\delta_h^l(t)}_j \frac{\partial}{\partial z_o^l(t)_j} \left(\boldsymbol{o_t^l} \circ \tanh(\boldsymbol{c_t^l})\right)_j && \text{(expression equals 0 except when } k = j\text{)} \\
&= \boldsymbol{\delta_h^l(t)}_j \operatorname{sigm}'(z_o^l(t))_j \tanh(\boldsymbol{c_t^l})_j && \text{(definition of } \boldsymbol{o_t^l} \text{ in terms of } z_o^l(t)\text{)}
\end{aligned}$$

Note that for the second equality we took the chain rule with respect to the elements of $\boldsymbol{h_t^l}$, because $\boldsymbol{o_t^l}$ affects $\boldsymbol{h_t^l}$. □

Lemma 6. *For $l \in \{1, \ldots, L\}$ and $t \in \{1, \ldots, T\}$,*

$$\boldsymbol{\delta_{\hat{h}}^l(t)} = \boldsymbol{\delta_c^l(t)} \circ \boldsymbol{i_t^l} \circ \tanh'(z_{\hat{h}}^l(t)) \tag{14}$$

Proof. We prove this element-wise. For any $j = 1, \ldots n$:

$$\begin{aligned}
\boldsymbol{\delta_{\hat{h}}^l(t)}_j &= \frac{\partial L}{\partial z_{\hat{h}}^l(t)_j} && \text{(definition of } \boldsymbol{\delta_{\hat{h}}^l(t)}\text{)} \\
&= \sum_{k=1}^{n} \frac{\partial L}{\partial(\boldsymbol{c_t^l})_k} \frac{\partial(\boldsymbol{c_t^l})_k}{\partial z_{\hat{h}}^l(t)_j} && \text{(chain rule)} \\
&= \sum_{k=1}^{n} \boldsymbol{\delta_c^l(t)}_k \frac{\partial}{\partial z_{\hat{h}}^l(t)_j} \left(\boldsymbol{f_t^l} \circ \boldsymbol{c_{t-1}^l} + \boldsymbol{i_t^l} \circ \boldsymbol{\hat{h}_t^l}\right)_k && \text{(definition of } \boldsymbol{\delta_c^l(t)} \text{ and } \boldsymbol{c_t^l}\text{)} \\
&= \boldsymbol{\delta_c^l(t)}_j \frac{\partial}{\partial z_{\hat{h}}^l(t)_j} \left(\boldsymbol{i_t^l} \circ \boldsymbol{\hat{h}_t^l}\right)_j && \text{(expression equals 0 except when } k = j\text{)} \\
&= \boldsymbol{\delta_c^l(t)}_j (\boldsymbol{i_t^l})_j \tanh'(z_{\hat{h}}^l(t))_j && \text{(definition of } \boldsymbol{\hat{h}_t^l} \text{ in terms of } z_{\hat{h}}^l(t)\text{)}
\end{aligned}$$

Note that for the second equality we took the chain rule with respect to the elements of $\boldsymbol{c_t^l}$, because $\boldsymbol{\hat{h}_t^l}$ affects $\boldsymbol{c_t^l}$. □

Lemma 7. *For all $l \in \{1, \ldots, L\}$,*

$$\frac{\partial L}{\partial U_i^l} = \sum_{t=1}^{T} (\boldsymbol{h_t^{l-1}})(\boldsymbol{\delta_i^l(t)})^\top \qquad \frac{\partial L}{\partial V_i^l} = \sum_{t=1}^{T} (\boldsymbol{h_{t-1}^l})(\boldsymbol{\delta_i^l(t)})^\top$$
$$\frac{\partial L}{\partial U_f^l} = \sum_{t=1}^{T} (\boldsymbol{h_t^{l-1}})(\boldsymbol{\delta_f^l(t)})^\top \qquad \frac{\partial L}{\partial V_f^l} = \sum_{t=1}^{T} (\boldsymbol{h_{t-1}^l})(\boldsymbol{\delta_f^l(t)})^\top$$
$$\frac{\partial L}{\partial U_o^l} = \sum_{t=1}^{T} (\boldsymbol{h_t^{l-1}})(\boldsymbol{\delta_o^l(t)})^\top \qquad \frac{\partial L}{\partial V_o^l} = \sum_{t=1}^{T} (\boldsymbol{h_{t-1}^l})(\boldsymbol{\delta_o^l(t)})^\top$$
$$\frac{\partial L}{\partial U_{\hat{h}}^l} = \sum_{t=1}^{T} (\boldsymbol{h_t^{l-1}})(\boldsymbol{\delta_{\hat{h}}^l(t)})^\top \qquad \frac{\partial L}{\partial V_{\hat{h}}^l} = \sum_{t=1}^{T} (\boldsymbol{h_{t-1}^l})(\boldsymbol{\delta_{\hat{h}}^l(t)})^\top$$

Proof. We will prove the identities for the input gate i only; the proofs for f, o and $\hat{h}$ are identical. First recall Definition 2 for the weighted input:

$$z_i^l(t) = U_i \boldsymbol{h_t^{l-1}} + V_i \boldsymbol{h_{t-1}^l}$$

Now, for any $j, k \in \{1, \ldots, n\}$, consider the effect of $(U_i^l)_{jk}$. It maps from the kth element of $\boldsymbol{h_t^{l-1}}$ to the jth element of $z_i^l(t)$, for all t. Therefore applying the chain rule we obtain

$$\begin{aligned} \frac{\partial L}{\partial (U_i^l)_{jk}} &= \sum_{t=1}^{T} \frac{\partial L}{\partial z_i^l(t)_j} \frac{\partial z_i^l(t)_j}{\partial (U_i^l)_{jk}} && \text{(chain rule)} \\ &= \sum_{t=1}^{T} \boldsymbol{\delta_i^l(t)}_j (\boldsymbol{h_t^{l-1}})_k && \text{(definition of } \boldsymbol{\delta_i^l(t)} \text{ and } z_i^l(t)\text{)} \end{aligned}$$

Therefore

$$\frac{\partial L}{\partial (U_i^l)} = \sum_{t=1}^{T} (\boldsymbol{h_t^{l-1}})(\boldsymbol{\delta_i^l(t)})^\top \tag{15}$$

The expression for $\partial L / \partial V_i^l$ is derived similarly, by noting that $(V_i^l)_{jk}$ maps from the kth element of $\boldsymbol{h_{t-1}^l}$ to the jth element of $z_i^l(t)$. □

Corollary 1. *For all $l \in \{1, \ldots, L\}$,*

$$\begin{bmatrix} \partial L/\partial U_i^l & \partial L/\partial V_i^l \\ \partial L/\partial U_f^l & \partial L/\partial V_f^l \\ \partial L/\partial U_o^l & \partial L/\partial V_o^l \\ \partial L/\partial U_{\hat{h}}^l & \partial L/\partial V_{\hat{h}}^l \end{bmatrix} = \sum_{t=1}^{T} \begin{bmatrix} \boldsymbol{\delta_i^l(t)} \\ \boldsymbol{\delta_f^l(t)} \\ \boldsymbol{\delta_o^l(t)} \\ \boldsymbol{\delta_{\hat{h}}^l(t)} \end{bmatrix} \begin{bmatrix} \boldsymbol{h_t^{l-1}} & \boldsymbol{h_{t-1}^l} \end{bmatrix} \tag{16}$$

Proof. This is simply a rearrangement of Lemma 7. □

3.3 Summary

Now we have derived all the necessary equations, we have an algorithm to calculate the necessary error values for each LSTM block, and thus calculate the derivative of the loss function with respect to our various weights.

For $l \in \{1, \dots, L-1\}$ and $t \in \{1, \dots, T\}$, we calculate $\boldsymbol{\delta}_{\boldsymbol{h}}^{\boldsymbol{l}}(\boldsymbol{t})$ as follows:

$$\boldsymbol{\delta}_{\boldsymbol{h}}^{\boldsymbol{l}}(\boldsymbol{t}) = \begin{bmatrix} U_i^\top & U_f^\top & U_o^\top & U_{\hat{h}}^\top & V_i^\top & V_f^\top & V_o^\top & V_{\hat{h}}^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{\delta}_{\boldsymbol{i}}^{\boldsymbol{l}}(\boldsymbol{t}+\mathbf{1}) \\ \boldsymbol{\delta}_{\boldsymbol{f}}^{\boldsymbol{l}}(\boldsymbol{t}+\mathbf{1}) \\ \boldsymbol{\delta}_{\boldsymbol{o}}^{\boldsymbol{l}}(\boldsymbol{t}+\mathbf{1}) \\ \boldsymbol{\delta}_{\hat{\boldsymbol{h}}}^{\boldsymbol{l}}(\boldsymbol{t}+\mathbf{1}) \\ \boldsymbol{\delta}_{\boldsymbol{i}}^{\boldsymbol{l}+\mathbf{1}}(\boldsymbol{t}) \\ \boldsymbol{\delta}_{\boldsymbol{f}}^{\boldsymbol{l}+\mathbf{1}}(\boldsymbol{t}) \\ \boldsymbol{\delta}_{\boldsymbol{o}}^{\boldsymbol{l}+\mathbf{1}}(\boldsymbol{t}) \\ \boldsymbol{\delta}_{\hat{\boldsymbol{h}}}^{\boldsymbol{l}+\mathbf{1}}(\boldsymbol{t}) \end{bmatrix}$$

$$\begin{aligned}
\boldsymbol{\delta}_{\boldsymbol{c}}^{\boldsymbol{l}}(\boldsymbol{t}) &= \boldsymbol{\delta}_{\boldsymbol{c}}^{\boldsymbol{l}}(\boldsymbol{t}+\mathbf{1}) \circ f_{t+1}^{l} + \boldsymbol{\delta}_{\boldsymbol{h}}^{\boldsymbol{l}}(\boldsymbol{t}) \circ \boldsymbol{o}_{\boldsymbol{t}}^{\boldsymbol{l}} \circ \tanh'(\boldsymbol{c}_{\boldsymbol{t}}^{\boldsymbol{l}}) \\
\boldsymbol{\delta}_{\boldsymbol{i}}^{\boldsymbol{l}}(\boldsymbol{t}) &= \boldsymbol{\delta}_{\boldsymbol{c}}^{\boldsymbol{l}}(\boldsymbol{t}) \circ \mathrm{sigm}'(z_i^l(t)) \circ \hat{\boldsymbol{h}}_{\boldsymbol{t}}^{\boldsymbol{l}} \\
\boldsymbol{\delta}_{\boldsymbol{o}}^{\boldsymbol{l}}(\boldsymbol{t}) &= \boldsymbol{\delta}_{\boldsymbol{h}}^{\boldsymbol{l}}(\boldsymbol{t}) \circ \mathrm{sigm}'(z_o^l(t)) \circ \tanh(\boldsymbol{c}_{\boldsymbol{t}}^{\boldsymbol{l}}) \\
\boldsymbol{\delta}_{\boldsymbol{f}}^{\boldsymbol{l}}(\boldsymbol{t}) &= \boldsymbol{\delta}_{\boldsymbol{c}}^{\boldsymbol{l}}(\boldsymbol{t}) \circ \mathrm{sigm}'(z_f^l(t)) \circ \boldsymbol{c}_{\boldsymbol{t}-\mathbf{1}}^{\boldsymbol{l}} \\
\boldsymbol{\delta}_{\hat{\boldsymbol{h}}}^{\boldsymbol{l}}(\boldsymbol{t}) &= \boldsymbol{\delta}_{\boldsymbol{c}}^{\boldsymbol{l}}(\boldsymbol{t}) \circ \boldsymbol{i}_{\boldsymbol{t}}^{\boldsymbol{l}} \circ \tanh'(z_{\hat{h}}^l(t))
\end{aligned}$$

Note: if $t = T$ then we take $\boldsymbol{\delta}^{\boldsymbol{l}}(\boldsymbol{t}+\mathbf{1})$ to be zero for $\boldsymbol{i}$, $\boldsymbol{f}$, $\boldsymbol{o}$, $\hat{\boldsymbol{h}}$ and $\boldsymbol{c}$. if $l = L$ how do we calculate $\boldsymbol{\delta}_{\boldsymbol{h}}^{\boldsymbol{l}}(\boldsymbol{t})$?

Once we have calculated the above error values for all l and t, we can calculate the derivative of the loss function with respect to our various weights. In particular, for $l \in \{1, \dots, L\}$:

$$\begin{bmatrix} \partial L/\partial U_i^l & \partial L/\partial V_i^l \\ \partial L/\partial U_f^l & \partial L/\partial V_f^l \\ \partial L/\partial U_o^l & \partial L/\partial V_o^l \\ \partial L/\partial U_{\hat{h}}^l & \partial L/\partial V_{\hat{h}}^l \end{bmatrix} = \sum_{t=1}^{T} \begin{bmatrix} \boldsymbol{\delta}_{\boldsymbol{i}}^{\boldsymbol{l}}(\boldsymbol{t}) \\ \boldsymbol{\delta}_{\boldsymbol{f}}^{\boldsymbol{l}}(\boldsymbol{t}) \\ \boldsymbol{\delta}_{\boldsymbol{o}}^{\boldsymbol{l}}(\boldsymbol{t}) \\ \boldsymbol{\delta}_{\hat{\boldsymbol{h}}}^{\boldsymbol{l}}(\boldsymbol{t}) \end{bmatrix} \begin{bmatrix} \boldsymbol{h}_{\boldsymbol{t}}^{\boldsymbol{l}-\mathbf{1}} & \boldsymbol{h}_{\boldsymbol{t}-\mathbf{1}}^{\boldsymbol{l}} \end{bmatrix}$$

We then use these derivatives to apply gradient descent to U^l and V^l.

4 Random

$\boldsymbol{\delta}_{\boldsymbol{c}^{(2)}}$

$\boldsymbol{\delta}_{\boldsymbol{h}^{(2)}}$

$\boldsymbol{\delta}_{\boldsymbol{c}^{(1)}}$

$\boldsymbol{\delta}_{\boldsymbol{h}^{(1)}}$

$\boldsymbol{\delta}_{\boldsymbol{c}} \mathrel{+}= \boldsymbol{\delta}_{\boldsymbol{h}} \boldsymbol{o}_{\boldsymbol{t}} \tanh'(\boldsymbol{c}_{\boldsymbol{t}})$

$\boldsymbol{\delta}_{\boldsymbol{c}} = \boldsymbol{\delta}_{\boldsymbol{c}} \circ \boldsymbol{f}_{\boldsymbol{t}}$

$\boldsymbol{\delta}_{\boldsymbol{h}} \mathrel{+}= \text{upper grad}$

5 Other Recurrent Units

Different recurrent units:

RNN

$$\boldsymbol{h_t} = \sigma\left(\boldsymbol{T}_{n\times 2n}\begin{bmatrix}\boldsymbol{x_t}\\ \boldsymbol{h_{t-1}}\end{bmatrix}\right) \tag{17}$$

$$\boldsymbol{T}_{n\times 2n} = [\boldsymbol{W_{xh}}\boldsymbol{W_{hh}}] \tag{18}$$

$$\boldsymbol{h}_t = \sigma\left(\boldsymbol{W_{xh}}\boldsymbol{x}_t + \boldsymbol{W_{hh}}\boldsymbol{h}_{t-1}\right) \tag{19}$$

$$\frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{h}_{t-1}} = \operatorname{diag}\left(\sigma'(\ldots)\right)\boldsymbol{W_{hh}}^{\top} \tag{20}$$

$$\begin{aligned}\left\|\frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{h}_{t-1}}\right\| &\leq \left\|\operatorname{diag}\left(\sigma'(\ldots)\right)\right\|\left\|\boldsymbol{W_{hh}}^{\top}\right\| &(21)\\ &\leq \gamma\lambda_1 &(22)\end{aligned}$$

$$\left\|\frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{h}_{t-k}}\right\| \leq \left(\gamma\lambda_1\right)^k \to 0 \quad \text{if } \lambda_1 < \frac{1}{\gamma} \tag{23}$$

$$\frac{\partial \boldsymbol{c}_t}{\partial \boldsymbol{c}_{t-1}} = \boldsymbol{I} \tag{24}$$

GRU [Cho et al., 2014]

$$\begin{aligned}\begin{pmatrix}\boldsymbol{z_t}\\ \boldsymbol{r_t}\end{pmatrix} &= \begin{pmatrix}\text{sigm}\\ \text{sigm}\end{pmatrix}\boldsymbol{T}_{2n\times 2n}\begin{bmatrix}\boldsymbol{x_t}\\ \boldsymbol{h_{t-1}}\end{bmatrix} &(25)\\ \boldsymbol{\hat{h}_t} &= \tanh(\boldsymbol{W}\boldsymbol{x_t} + \boldsymbol{r_t}\circ\boldsymbol{U}\boldsymbol{h_{t-1}}) &(26)\\ \boldsymbol{h_t} &= \boldsymbol{z_t}\circ\boldsymbol{h_{t-1}} + (1-\boldsymbol{z_t})\circ\boldsymbol{\hat{h}_t} &(27)\end{aligned}$$

My unit (maybe we should try to implement this!)

$$\begin{aligned}\begin{pmatrix}\boldsymbol{i_t}\\ \boldsymbol{f_t}\\ \boldsymbol{\hat{h}_t}\end{pmatrix} &= \begin{pmatrix}\text{sigm}\\ \text{sigm}\\ \text{tanh}\end{pmatrix}\boldsymbol{T}_{3n\times 2n}\begin{bmatrix}\boldsymbol{x_t}\\ \boldsymbol{h_{t-1}}\end{bmatrix} &(28)\\ \boldsymbol{h_t} &= \boldsymbol{f_t}\circ\boldsymbol{h_{t-1}} + \boldsymbol{i_t}\circ\boldsymbol{\hat{h}_t} &(29)\end{aligned}$$

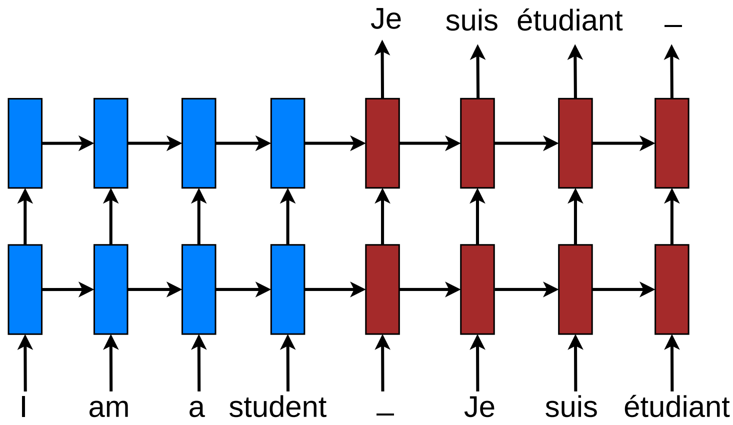

Figure 1: **NMT**

6 Neural Machine Translation

[Sutskever et al., 2014]

6.1 Attention

Content-based

$$\boldsymbol{a}_t = \mathit{Attend}(\boldsymbol{h}_{t-1}, \bar{\boldsymbol{h}}_{1\ldots S}) \tag{30}$$

Location-based

$$\boldsymbol{a}_t = \mathit{Attend}(\boldsymbol{h}_{t-1}, \boldsymbol{a}_{t-1}) \tag{31}$$

Hybrid

$$\boldsymbol{a}_t = \mathit{Attend}(\boldsymbol{h}_{t-1}, \boldsymbol{a}_{t-1}, \bar{\boldsymbol{h}}_{1\ldots S}) \tag{32}$$

7 Conclusion and Future Work

References

[Cho et al., 2014] Cho, K., van Merrienboer, B., Gulcehre, C., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP*.

[Sutskever et al., 2014] Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *NIPS*.

[Zaremba et al., 2014] Zaremba, W., Sutskever, I., and Vinyals, O. (2014). Recurrent neural network regularization. *CoRR*, abs/1409.2329.