DL Assignment 3

Sanchit Agrawal - CS13B061
Ishu Garg - CS13B060

April 2017

1 LSTM Equations

The LSTM equations are shown below for reference.

1.1 Gate Equations

$$o_t = \sigma(W_o h_{t-1} + U_o x_t + b_o) \tag{1}$$
$$i_t = \sigma(W_i h_{t-1} + U_i x_t + b_i) \tag{2}$$
$$f_t = \sigma(W_f h_{t-1} + U_f x_t + b_f) \tag{3}$$

1.2 State Equations

$$\widetilde{s_t} = \sigma(W h_{t-1} + U x_t + b) \tag{4}$$
$$s_t = f_t \odot s_{t-1} + i_t \odot \widetilde{s_t} \tag{5}$$
$$h_t = o_t \odot \sigma(s_t) \tag{6}$$

h_t and L_t are respectively the outputs and loss at time-step t of the LSTM.

2 Notation

We introduce some notation to make the derivatives simpler to express.

- Let $x \in \mathbb{R}^n$. Then $d_{r-1}(x)$ denotes the diagonal tensor in $\mathbb{R}^{n^r}$ with x as its diagonal entries and r is the rank of the tensor. $d(x)$ (without a subscript) denotes $d_1(x)$. Specifically, $d(x)$ is a diagonal matrix with diagonal x.
- Let $x \in \mathbb{R}^n$, f is a real valued function and $f(x)$ be a vector with $f(x)_i = f(x_i)$ (f is broadcasted over x). Then $f'(x)$ denotes a vector where $f'(x)_i = \frac{\partial f(x_i)}{\partial x_i}$.
- Let g_t denote any of $\{o_t, i_t, f_t, \widetilde{s_t}\}$.

- Let A_g denote any parameter (matrix or vector) W or U or b in the equations of g_t.
- Let $A_{p \neq g}$ denotes any parameter (matrix or vector) not belonging to equations of the g_t.

3 Generic Derivative Computations

3.1 $\frac{\partial s_t}{\partial A}$ (A is arbitrary)

Let A be any quantity with respect to which the derivative is required (typically a parameter, gate, or state). Then the following equation holds:

$$\begin{aligned}\frac{\partial s_t}{\partial A} &= \frac{\partial(f_t \odot s_{t-1} + i_t \odot \widetilde{s}_t)}{\partial A} \\ &= \frac{\partial f_t \odot s_{t-1}}{\partial f_t}\frac{\partial f_t}{\partial A} + \frac{\partial f_t \odot s_{t-1}}{\partial s_{t-1}}\frac{\partial s_{t-1}}{\partial A} + \frac{\partial i_t \odot \widetilde{s}_t}{\partial \widetilde{s}_t}\frac{\widetilde{s}_t}{\partial A} + \frac{\partial i_t \odot \widetilde{s}_t}{\partial i_t}\frac{\partial i_t}{\partial A} \\ &= d(s_{t-1})\frac{\partial f_t}{\partial A} + d(f_t)\frac{\partial s_{t-1}}{\partial A} + d(i_t)\frac{\partial \widetilde{s}_t}{\partial A} + d(\widetilde{s}_t)\frac{\partial i_t}{\partial A}\end{aligned} \tag{7}$$

3.2 $\frac{\partial h_t}{\partial A}$ (A is arbitrary)

Let A be any quantity with respect to which the derivative is required (typically a parameter, gate, or state). Then the following equation holds:

$$\begin{aligned}\frac{\partial h_t}{\partial A} &= \frac{\partial h_t}{\partial o_t}\frac{\partial o_t}{\partial A} + \frac{\partial h_t}{\partial \sigma(s_t)}\frac{\partial \sigma(s_t)}{\partial s_t}\frac{\partial s_t}{\partial A} \\ &= d(\sigma(s_t))\frac{\partial o_t}{\partial A} + d(o_t)d(\sigma'(s_t))\frac{\partial s_t}{\partial A} \\ &= d(\sigma(s_t))\frac{\partial o_t}{\partial A} \\ &+ d(o_t)d(\sigma'(s_t))\left(d(s_{t-1})\frac{\partial f_t}{\partial A} + d(f_t)\frac{\partial s_{t-1}}{\partial A} + d(i_t)\frac{\widetilde{s}_t}{\partial A} + d(\widetilde{s}_t)\frac{\partial i_t}{\partial A}\right)\end{aligned} \tag{8}$$

3.3 $\frac{\partial L_t}{\partial A}$ (A is arbitrary)

Let A be any quantity with respect to which the derivative is required (typically a parameter, gate, or state). Then the following equation holds:

$$
\begin{aligned}
\frac{\partial L_t}{\partial A} &= \frac{\partial L_t}{\partial h_t}\frac{\partial h_t}{\partial A} = \frac{\partial L_t}{\partial h_t}\left(d(\sigma(s_t))\frac{\partial o_t}{\partial A} + d(o_t)d(\sigma'(s_t))\frac{\partial s_t}{\partial A}\right) \\
&= \frac{\partial L_t}{\partial h_t}\Bigg(d(\sigma(s_t))\frac{\partial o_t}{\partial A} + \\
& d(o_t)d(\sigma'(s_t))\left(d(s_{t-1})\frac{\partial f_t}{\partial A} + d(f_t)\frac{\partial s_{t-1}}{\partial A} + d(i_t)\frac{\widetilde{s_t}}{\partial A} + d(\widetilde{s_t})\frac{\partial i_t}{\partial A}\right)\Bigg)
\end{aligned} \tag{9}
$$

3.4 $\frac{\partial g_t}{A_{p \neq g}}$ $\left(\frac{\partial g_t}{\partial W_{p \neq g}}\right.$ or $\frac{\partial g_t}{\partial U_{p \neq g}}$ or $\left.\frac{\partial g_t}{\partial b_{p \neq g}}\right)$

$$
\begin{aligned}
\frac{\partial g_t}{\partial A} &= \frac{\partial \sigma(W_g h_{t-1} + U_g x_t + b_g)}{\partial (W_g h_{t-1} + U_g x_t + b_g)} \frac{\partial (W_g h_{t-1} + U_g x_t + b_g)}{\partial A} \\
&= d(\sigma'(W_g h_{t-1} + U_g x_t + b_g)) W_g \frac{\partial h_{t-1}}{\partial A} \\
&= d(g_t') W_g \frac{\partial h_{t-1}}{\partial A}
\end{aligned} \tag{10}
$$

3.5 $\frac{\partial g_t}{\partial W_g}$

Let g_t denote any of $\{o_t, i_t, f_t, \widetilde{s_t}\}$, and W_g denote the W weight matrix belonging to g_t. Then the following equation holds:

$$
\begin{aligned}
\frac{\partial g_t}{\partial W_g} &= \frac{\partial \sigma(W_g h_{t-1} + U_g x_t + b_g)}{\partial (W_g h_{t-1} + U_g x_t + b_g)} \frac{\partial (W_g h_{t-1} + U_g x_t + b_g)}{\partial W_g} \\
&= d(\sigma'(W_g h_{t-1} + U_g x_t + b_g)) \frac{\partial W_g h_{t-1}}{\partial W_g} \\
&= d(g_t') \left(\frac{\partial W_g}{\partial W_g} h_{t-1} + W_g \frac{\partial h_{t-1}}{\partial W_g} \right) \\
&= d(g_t') \left(d_2(h_{t-1}) + W_g \frac{\partial h_{t-1}}{\partial W_g} \right)
\end{aligned} \tag{11}
$$

3.6 $\frac{\partial g_t}{\partial U_g}$

Let g_t denote any of $\{o_t, i_t, f_t, \widetilde{s_t}\}$, and U_g denote the U weight matrix belonging to g_t. Then the following equation holds:

$$
\begin{aligned}
\frac{\partial g_t}{\partial U_g} &= \frac{\partial \sigma(W_g h_{t-1} + U_g x_t + b_g)}{\partial(W_g h_{t-1} + U_g x_t + b_g)} \frac{\partial(W_o h_{t-1} + U_o x_t + b_o)}{\partial U_o} \\
&= d(g_t') \left(W_g \frac{\partial h_{t-1}}{\partial U_g} + \frac{\partial U_g x_t}{\partial U_g} \right) \\
&= d(g_t') \left(W_g \frac{\partial h_{t-1}}{\partial U_g} + d_2(x_t) \right)
\end{aligned} \tag{12}
$$

3.7 $\frac{\partial g_t}{\partial b_g}$

Let g_t denote any of $\{o_t, i_t, f_t, \widetilde{s_t}\}$, and b_g denote the b bias vector belonging to g_t. Then the following equation holds:

$$
\begin{aligned}
\frac{\partial g_t}{\partial b_g} &= \frac{\partial \sigma(W_g h_{t-1} + U_g x_t + b_g)}{\partial(W_g h_{t-1} + U_g x_t + b_g)} \frac{\partial(W_g h_{t-1} + U_g x_t + b_g)}{\partial b_g} \\
&= d(g_t') \left(W_g \frac{\partial h_{t-1}}{\partial b_g} + I \right)
\end{aligned} \tag{13}
$$

4 Specific Derivative Computations

We compute $\frac{\partial h_t}{\partial A}$ for all the parameters A. Also since $\frac{\partial L_t}{\partial A} = \frac{\partial L_t}{\partial h_t} \frac{\partial h_t}{\partial A}$ for all the parameters A, we only show explicit computations for $\frac{\partial h_t}{\partial A}$.

4.1 $\frac{\partial h_t}{\partial W}$

$$\begin{aligned}
\frac{\partial h_t}{\partial W} &= d(\sigma(s_t))\frac{\partial o_t}{\partial W}+ \\
&d(o_t)d(\sigma'(s_t))\left(d(s_{t-1})\frac{\partial f_t}{\partial W}+d(f_t)\frac{\partial s_{t-1}}{\partial W}+d(i_t)\frac{\widetilde{s}_t}{\partial W}+d(\widetilde{s}_t)\frac{\partial i_t}{\partial W}\right) \\
&= d(\sigma(s_t))d(o_t')W_o\frac{\partial h_{t-1}}{\partial W}+ \\
&d(o_t)d(\sigma'(s_t))\left(d(s_{t-1})d(f_t')W_f\frac{\partial h_{t-1}}{\partial W}+d(f_t)\frac{\partial s_{t-1}}{\partial W}+\right. \\
&\left.d(i_t)\left(d(\widetilde{s}_t{}')\left(d_2(h_{t-1})+W\frac{\partial h_{t-1}}{\partial W}\right)\right)+d(\widetilde{s}_t)\left(d(i_t')W_i\frac{\partial h_{t-1}}{\partial W}\right)\right) \\
&= \left(d(\sigma(s_t))d(o_t')W_o+d(o_t)d(\sigma'(s_t))(d(s_{t-1})d(f_t')W_f+d(i_t)d(\widetilde{s}_t{}')W+d(\widetilde{s}_t)d(i_t')W_i)\right)\frac{\partial h_{t-1}}{\partial W} \\
&+d(o_t)d(\sigma'(s_t))d(f_t)\frac{\partial s_{t-1}}{\partial W}+d(o_t)d(\sigma'(s_t))d(i_t)d(\widetilde{s}_t{}')d_2(h_{t-1})
\end{aligned} \tag{14}$$

4.2 $\frac{\partial h_t}{\partial W_i}$

$$\begin{aligned}
\frac{\partial h_t}{\partial W_i} &= d(\sigma(s_t))\frac{\partial o_t}{\partial W_i} \\
&+d(o_t)d(\sigma'(s_t))\left(d(s_{t-1})\frac{\partial f_t}{\partial W_i}+d(f_t)\frac{\partial s_{t-1}}{\partial W_i}+d(i_t)\frac{\widetilde{s}_t}{\partial W_i}+d(\widetilde{s}_t)\frac{\partial i_t}{\partial W_i}\right) \\
&= d(\sigma(s_t))d(o_t')W_o\frac{\partial h_{t-1}}{\partial W_i}+d(o_t)d(\sigma'(s_t))\left(d(s_{t-1})d(f_t')W_f\frac{\partial h_{t-1}}{\partial W_i}\right. \\
&\left.+d(f_t)\frac{\partial s_{t-1}}{\partial W_i}+d(i_t)d(\widetilde{s}_t{}^{'})W\frac{\partial h_{t-1}}{\partial W_i}+d(\widetilde{s}_t)d(i_t')\left(d_2(h_{t-1})+W_i\frac{\partial h_{t-1}}{\partial W_i}\right)\right) \\
&= \left(d(\sigma(s_t))d(o_t')W_o+d(o_t)d(\sigma'(s_t))\Big(d(s_{t-1})d(f_t')W_f+d(i_t)d(\widetilde{s}_t{}^{'})W+d(\widetilde{s}_t)d(i_t')W_i\Big)\right)\frac{\partial h_{t-1}}{\partial W_i} \\
&+d(o_t)d(\sigma'(s_t))d(f_t)\frac{\partial s_{t-1}}{\partial W_i}+d(o_t)d(\sigma'(s_t))d(\widetilde{s}_t)d(i_t')d_2(h_{t-1})
\end{aligned} \tag{15}$$

4.3 $\frac{\partial h_t}{\partial W_f}$

$$
\begin{aligned}
\frac{\partial h_t}{\partial W_f} &= d(\sigma(s_t))\frac{\partial o_t}{\partial W_f} \\
&+ d(o_t)d(\sigma'(s_t))\left(d(s_{t-1})\frac{\partial f_t}{\partial W_f} + d(f_t)\frac{\partial s_{t-1}}{\partial W_f} + d(i_t)\frac{\widetilde{s}_t}{\partial W_f} + d(\widetilde{s}_t)\frac{\partial i_t}{\partial W_f}\right) \\
&= d(\sigma(s_t))d(o_t')W_o\frac{\partial h_{t-1}}{\partial W_f} + d(o_t)d(\sigma'(s_t))d(s_{t-1})d(f_t')\Big(W_f\frac{\partial h_{t-1}}{\partial W_f}d_2(h_{t-1})\Big) \\
&+ d(o_t)d(\sigma'(s_t))\left(d(f_t)\frac{\partial s_{t-1}}{\partial W_f} + d(i_t)d(\widetilde{s_t}^{'})W\frac{\partial h_{t-1}}{\partial W_f} + d(\widetilde{s}_t)d(i_t')W_i\frac{\partial h_{t-1}}{\partial W_f}\right) \\
&= \left(d(\sigma(s_t))d(o_t')W_o + d(o_t)d(\sigma'(s_t))\Big(d(s_{t-1})d(f_t')W_f + d(i_t)d(\widetilde{s_t}^{'})W + d(\widetilde{s}_t)d(i_t')W_i\Big)\right)\frac{\partial h_{t-1}}{\partial W_f} \\
&+ d(o_t)d(\sigma'(s_t))d(f_t)\frac{\partial s_{t-1}}{\partial W_f} + d(o_t)d(\sigma'(s_t))d(s_{t-1})d(f_t')d_2(h_{t-1})
\end{aligned} \tag{16}
$$

4.4 $\frac{\partial h_t}{\partial W_o}$

$$
\begin{aligned}
\frac{\partial h_t}{\partial W_o} &= d(\sigma(s_t))\frac{\partial o_t}{\partial W_o} \\
&+ d(o_t)d(\sigma'(s_t))\left(d(s_{t-1})\frac{\partial f_t}{\partial W_o} + d(f_t)\frac{\partial s_{t-1}}{\partial W_o} + d(i_t)\frac{\widetilde{s}_t}{\partial W_o} + d(\widetilde{s}_t)\frac{\partial i_t}{\partial W_o}\right) \\
&= d(\sigma(s_t))d(o_t')\left(d_2(h_{t-1}) + W_o\frac{\partial h_{t-1}}{\partial W_o}\right) \\
&+ d(o_t)d(\sigma'(s_t))\left(d(s_{t-1})d(f_t')W_f\frac{\partial h_{t-1}}{\partial W_o}d(f_t)\frac{\partial s_{t-1}}{\partial W_o} + d(i_t)d(\widetilde{s_t}^{'})W\frac{\partial h_{t-1}}{\partial W_o} + d(\widetilde{s}_t)d(i_t')W_i\frac{\partial h_{t-1}}{\partial W_o}\right) \\
&= \left(d(\sigma(s_t))d(o_t')W_o + d(o_t)d(\sigma'(s_t))\Big(d(s_{t-1})d(f_t')W_f + d(i_t)d(\widetilde{s_t}^{'})W + d(\widetilde{s}_t)d(i_t')W_i\Big)\right)\frac{\partial h_{t-1}}{\partial W_o} \\
&+ d(o_t)d(\sigma'(s_t))d(f_t)\frac{\partial s_{t-1}}{\partial W_o} + d(s_t)d(o_t')d_2(h_{t-1})
\end{aligned} \tag{17}
$$

4.5 $\frac{\partial h_t}{\partial U}$

$$
\begin{aligned}
\frac{\partial h_t}{\partial U} &= d(\sigma(s_t))\frac{\partial o_t}{\partial U} + \\
&d(o_t)d(\sigma'(s_t))\Bigg(d(s_{t-1})\frac{\partial f_t}{\partial U} + d(f_t)\frac{\partial s_{t-1}}{\partial U} + d(i_t)\frac{\widetilde{s}_t}{\partial U} + d(\widetilde{s}_t)\frac{\partial i_t}{\partial U}\Bigg) \\
&= d(\sigma(s_t))d(o_t')W_o\frac{\partial h_{t-1}}{\partial U} + \\
&d(o_t)d(\sigma'(s_t))\Bigg(d(s_{t-1})d(f_t')W_f\frac{\partial h_{t-1}}{\partial U} + d(f_t)\frac{\partial s_{t-1}}{\partial U} + \\
&d(i_t)d(\widetilde{s}_t{'})\Bigg(d_2(x_t) + W\frac{\partial h_{t-1}}{\partial U}\Bigg) + d(\widetilde{s}_t)d(i_t')W_i\frac{\partial h_{t-1}}{\partial U}\Bigg) \\
&= \Bigg(d(\sigma(s_t))d(o_t')W_o + d(o_t)d(\sigma'(s_t))\Big(d(s_{t-1})d(f_t')W_f + d(i_t)d(\widetilde{s}_t{'})W + d(\widetilde{s}_t)d(i_t')W_i\Big)\Bigg)\frac{\partial h_{t-1}}{\partial U} \\
&+ d(o_t)d(\sigma'(s_t))d(f_t)\frac{\partial s_{t-1}}{\partial U} + d(o_t)d(\sigma'(s_t))d(i_t)d(\widetilde{s}_t{'})d_2(x_t)
\end{aligned}
\tag{18}
$$

4.6 $\frac{\partial h_t}{\partial U_i}$

$$
\begin{aligned}
\frac{\partial h_t}{\partial U_i} &= d(\sigma(s_t))\frac{\partial o_t}{\partial U_i} \\
&+ d(o_t)d(\sigma'(s_t))\Bigg(d(s_{t-1})\frac{\partial f_t}{\partial U_i} + d(f_t)\frac{\partial s_{t-1}}{\partial U_i} + d(i_t)\frac{\widetilde{s}_t}{\partial U_i} + d(\widetilde{s}_t)\frac{\partial i_t}{\partial U_i}\Bigg) \\
&= d(\sigma(s_t))d(o_t')W_o\frac{\partial h_{t-1}}{\partial U_i} + d(o_t)d(\sigma'(s_t))\Bigg(d(s_{t-1})d(f_t')W_f\frac{\partial h_{t-1}}{\partial U_i} \\
&+ d(f_t)\frac{\partial s_{t-1}}{\partial U_i} + d(i_t)d(\widetilde{s}_t^{'})W\frac{\partial h_{t-1}}{\partial U_i} + d(\widetilde{s}_t)d(i_t')\Bigg(d_2(x_t) + W_i\frac{\partial h_{t-1}}{\partial U_i}\Bigg)\Bigg) \\
&= \Bigg(d(\sigma(s_t))d(o_t')W_o + d(o_t)d(\sigma'(s_t))\Big(d(s_{t-1})d(f_t')W_f + d(i_t)d(\widetilde{s}_t^{'})W + d(\widetilde{s}_t)d(i_t')W_i\Big)\Bigg)\frac{\partial h_{t-1}}{\partial U_i} \\
&+ d(o_t)d(\sigma'(s_t))d(f_t)\frac{\partial s_{t-1}}{\partial U_i} + d(o_t)d(\sigma'(s_t))d(\widetilde{s}_t)d(i_t')d_2(x_t)
\end{aligned}
\tag{19}
$$

4.7 $\frac{\partial h_t}{\partial U_f}$

$$\begin{aligned}
\frac{\partial h_t}{\partial U_f} &= d(\sigma(s_t))\frac{\partial o_t}{\partial U_f} \\
&+ d(o_t)d(\sigma'(s_t))\left(d(s_{t-1})\frac{\partial f_t}{\partial U_f} + d(f_t)\frac{\partial s_{t-1}}{\partial U_f} + d(i_t)\frac{\widetilde{s}_t}{\partial U_f} + d(\widetilde{s}_t)\frac{\partial i_t}{\partial U_f}\right) \\
&= d(\sigma(s_t))d(o_t')W_o\frac{\partial h_{t-1}}{\partial U_f} + d(o_t)d(\sigma'(s_t))d(s_{t-1})d(f_t')\Big(W_f\frac{\partial h_{t-1}}{\partial U_f}d_2(x_t)\Big) \\
&+ d(o_t)d(\sigma'(s_t))\left(d(f_t)\frac{\partial s_{t-1}}{\partial U_f} + d(i_t)d(\widetilde{s_t}^{\,'})W\frac{\partial h_{t-1}}{\partial U_f} + d(\widetilde{s}_t)d(i_t')W_i\frac{\partial h_{t-1}}{\partial U_f}\right) \\
&= \left(d(\sigma(s_t))d(o_t')W_o + d(o_t)d(\sigma'(s_t))\Big(d(s_{t-1})d(f_t')W_f + d(i_t)d(\widetilde{s_t}^{\,'})W + d(\widetilde{s}_t)d(i_t')W_i\Big)\right)\frac{\partial h_{t-1}}{\partial U_f} \\
&+ d(o_t)d(\sigma'(s_t))d(f_t)\frac{\partial s_{t-1}}{\partial U_f} + d(o_t)d(\sigma'(s_t))d(s_{t-1})d(f_t')d_2(x_t)
\end{aligned} \tag{20}$$

4.8 $\frac{\partial h_t}{\partial U_o}$

$$\begin{aligned}
\frac{\partial h_t}{\partial U_o} &= d(\sigma(s_t))\frac{\partial o_t}{\partial U_o} \\
&+ d(o_t)d(\sigma'(s_t))\left(d(s_{t-1})\frac{\partial f_t}{\partial U_o} + d(f_t)\frac{\partial s_{t-1}}{\partial U_o} + d(i_t)\frac{\widetilde{s}_t}{\partial U_o} + d(\widetilde{s}_t)\frac{\partial i_t}{\partial U_o}\right) \\
&= d(\sigma(s_t))d(o_t')\left(d_2(x_t) + W_o\frac{\partial h_{t-1}}{\partial U_o}\right) \\
&+ d(o_t)d(\sigma'(s_t))\left(d(s_{t-1})d(f_t')W_f\frac{\partial h_{t-1}}{\partial U_o}d(f_t)\frac{\partial s_{t-1}}{\partial U_o} + d(i_t)d(\widetilde{s_t}^{\,'})W\frac{\partial h_{t-1}}{\partial U_o} + d(\widetilde{s}_t)d(i_t')W_i\frac{\partial h_{t-1}}{\partial U_o}\right) \\
&= \left(d(\sigma(s_t))d(o_t')W_o + d(o_t)d(\sigma'(s_t))\Big(d(s_{t-1})d(f_t')W_f + d(i_t)d(\widetilde{s_t}^{\,'})W + d(\widetilde{s}_t)d(i_t')W_i\Big)\right)\frac{\partial h_{t-1}}{\partial U_o} \\
&+ d(o_t)d(\sigma'(s_t))d(f_t)\frac{\partial s_{t-1}}{\partial U_o} + d(s_t)d(o_t')d_2(x_t)
\end{aligned} \tag{21}$$

4.9 $\frac{\partial h_t}{\partial b}$

$$
\begin{aligned}
\frac{\partial h_t}{\partial b} &= d(\sigma(s_t))\frac{\partial o_t}{\partial b}+ \\
&d(o_t)d(\sigma'(s_t))\left(d(s_{t-1})\frac{\partial f_t}{\partial b}+d(f_t)\frac{\partial s_{t-1}}{\partial b}+d(i_t)\frac{\widetilde{s_t}}{\partial b}+d(\widetilde{s_t})\frac{\partial i_t}{\partial b}\right) \\
&= d(\sigma(s_t))d(o_t')W_o\frac{\partial h_{t-1}}{\partial b}+ \\
&d(o_t)d(\sigma'(s_t))\left(d(s_{t-1})d(f_t')W_f\frac{\partial h_{t-1}}{\partial b}+d(f_t)\frac{\partial s_{t-1}}{\partial b}+\right. \\
&\left.d(i_t)d(\widetilde{s_t}')\left(I+W\frac{\partial h_{t-1}}{\partial b}\right)+d(\widetilde{s_t})d(i_t')W_i\frac{\partial h_{t-1}}{\partial b}\right) \\
&= \left(d(\sigma(s_t))d(o_t')W_o+d(o_t)d(\sigma'(s_t))\Big(d(s_{t-1})d(f_t')W_f+d(i_t)d(\widetilde{s_t}')W+d(\widetilde{s_t})d(i_t')W_i\Big)\right)\frac{\partial h_{t-1}}{\partial b} \\
&+d(o_t)d(\sigma'(s_t))d(f_t)\frac{\partial s_{t-1}}{\partial b}+d(o_t)d(\sigma'(s_t))d(i_t)d(\widetilde{s_t}')
\end{aligned}
\tag{22}
$$

4.10 $\frac{\partial h_t}{\partial b_i}$

$$
\begin{aligned}
\frac{\partial h_t}{\partial b_i} &= d(\sigma(s_t))\frac{\partial o_t}{\partial b_i} \\
&+d(o_t)d(\sigma'(s_t))\left(d(s_{t-1})\frac{\partial f_t}{\partial b_i}+d(f_t)\frac{\partial s_{t-1}}{\partial b_i}+d(i_t)\frac{\widetilde{s_t}}{\partial b_i}+d(\widetilde{s_t})\frac{\partial i_t}{\partial b_i}\right) \\
&= d(\sigma(s_t))d(o_t')W_o\frac{\partial h_{t-1}}{\partial b_i}+d(o_t)d(\sigma'(s_t))\left(d(s_{t-1})d(f_t')W_f\frac{\partial h_{t-1}}{\partial b_i}\right. \\
&\left.+d(f_t)\frac{\partial s_{t-1}}{\partial b_i}+d(i_t)d(\widetilde{s_t}')W\frac{\partial h_{t-1}}{\partial b_i}+d(\widetilde{s_t})d(i_t')\left(I+W_i\frac{\partial h_{t-1}}{\partial b_i}\right)\right) \\
&= \left(d(\sigma(s_t))d(o_t')W_o+d(o_t)d(\sigma'(s_t))\Big(d(s_{t-1})d(f_t')W_f+d(i_t)d(\widetilde{s_t}')W+d(\widetilde{s_t})d(i_t')W_i\Big)\right)\frac{\partial h_{t-1}}{\partial b_i} \\
&+d(o_t)d(\sigma'(s_t))d(f_t)\frac{\partial s_{t-1}}{\partial b_i}+d(o_t)d(\sigma'(s_t))d(\widetilde{s_t})d(i_t')
\end{aligned}
\tag{23}
$$

4.11 $\frac{\partial h_t}{\partial b_f}$

$$\begin{aligned}
\frac{\partial h_t}{\partial b_f} &= d(\sigma(s_t))\frac{\partial o_t}{\partial b_f} \\
&+ d(o_t)d(\sigma'(s_t))\left(d(s_{t-1})\frac{\partial f_t}{\partial b_f} + d(f_t)\frac{\partial s_{t-1}}{\partial b_f} + d(i_t)\frac{\widetilde{s}_t}{\partial b_f} + d(\widetilde{s}_t)\frac{\partial i_t}{\partial b_f}\right) \\
&= d(\sigma(s_t))d(o_t')W_o\frac{\partial h_{t-1}}{\partial b_f} + d(o_t)d(\sigma'(s_t)) \\
&+ d(o_t)d(\sigma'(s_t))\left(d(f_t)\frac{\partial s_{t-1}}{\partial b_f} + d(i_t)d(\widetilde{s_t}')W\frac{\partial h_{t-1}}{\partial b_f} + d(\widetilde{s}_t)d(i_t')W_i\frac{\partial h_{t-1}}{\partial b_f}\right) \\
&= \left(d(\sigma(s_t))d(o_t')W_o + d(o_t)d(\sigma'(s_t))\Big(d(s_{t-1})d(f_t')W_f + d(i_t)d(\widetilde{s_t}')W + d(\widetilde{s}_t)d(i_t')W_i\Big)\right)\frac{\partial h_{t-1}}{\partial b_f} \\
&+ d(o_t)d(\sigma'(s_t))d(f_t)\frac{\partial s_{t-1}}{\partial b_f} + d(o_t)d(\sigma'(s_t))d(s_{t-1})d(f_t')
\end{aligned} \tag{24}$$

4.12 $\frac{\partial h_t}{\partial b_o}$

$$\begin{aligned}
\frac{\partial h_t}{\partial b_o} &= d(\sigma(s_t))\frac{\partial o_t}{\partial b_o} \\
&+ d(o_t)d(\sigma'(s_t))\left(d(s_{t-1})\frac{\partial f_t}{\partial b_o} + d(f_t)\frac{\partial s_{t-1}}{\partial b_o} + d(i_t)\frac{\widetilde{s}_t}{\partial b_o} + d(\widetilde{s}_t)\frac{\partial i_t}{\partial b_o}\right) \\
&= d(\sigma(s_t))d(o_t')\left(I + W_o\frac{\partial h_{t-1}}{\partial b_o}\right) \\
&+ d(o_t)d(\sigma'(s_t))\left(d(s_{t-1})d(f_t')W_f\frac{\partial h_{t-1}}{\partial b_o}d(f_t)\frac{\partial s_{t-1}}{\partial b_o} + d(i_t)d(\widetilde{s_t}')W\frac{\partial h_{t-1}}{\partial b_o} + d(\widetilde{s}_t)d(i_t')W_i\frac{\partial h_{t-1}}{\partial b_o}\right) \\
&= \left(d(\sigma(s_t))d(o_t')W_o + d(o_t)d(\sigma'(s_t))\Big(d(s_{t-1})d(f_t')W_f + d(i_t)d(\widetilde{s_t}')W + d(\widetilde{s}_t)d(i_t')W_i\Big)\right)\frac{\partial h_{t-1}}{\partial b_o} \\
&+ d(o_t)d(\sigma'(s_t))d(f_t)\frac{\partial s_{t-1}}{\partial b_o} + d(s_t)d(o_t')
\end{aligned} \tag{25}$$

5 LSTMs solve the Vanishing Gradient Problem

5.1 Problem Description

In RNNs, when the gradient of L_t w.r.t any parameter A is backpropagated through time, the norm of the gradient becomes smaller and smaller, until it vanishes for all practical purposes. Due to this, the earlier time steps are not held accountable for the current error in prediction. This is a serious problem since it prevents RNNs from learning long range temporal dependencies.

5.2 Intuitive Sketch of the Proof

If we are able to find any single path that allows the gradient to flow back in time without causing it to shrink, then even if the other gradient flow paths vanish, this particular path will prevent the overall gradient from vanishing. We will show the existence of such a common path in all the derivatives we have computed. This path is unique to the LSTMs, and is commonly known as the Constant Error Carousel in the LSTM literature.
If we observe the derivative $\frac{\partial L_t}{\partial A}$ for any parameter A, then it can be written in the following form:

$$\frac{\partial h_t}{\partial A} = M_t \frac{h_{t-1}}{\partial A} + N_t \frac{\partial s_{t-1}}{\partial A} + V_t \tag{26}$$

Here M_t, N_t, V_t are tensors that are functions of the gates and states at time t. In the next section, we will show that the 2^{nd} term $N_t \frac{\partial s_{t-1}}{\partial A}$ doesn't vanish even when it is unrolled to a large number of time steps.

5.3 Proof

Consider the 2^{nd} term (represented as $N_t \frac{\partial s_{t-1}}{\partial A}$ in equation 26) in each of the derivatives. Represent this term by T_0.

$$\begin{aligned} T_0 &= d(o_t)d(\sigma'(s_t))d(f_t)\frac{\partial s_{t-1}}{\partial A} \\ &= d(o_t)d(\sigma'(s_t))d(f_t)\Big(d(s_{t-2})\frac{\partial f_{t-1}}{\partial A} + d(f_{t-1})\frac{\partial s_{t-2}}{\partial A} + d(i_{t-1})\frac{\partial \widetilde{s}_{t-1}}{\partial A} + d(\widetilde{s}_{t-1})\frac{\partial i_{t-1}}{\partial A}\Big) \end{aligned} \tag{27}$$

Consider the 2^{nd} term of equation 27. Call this term T_1. We will show that this term by itself provides a path for the gradient to flow back in time. In other words, even if the other terms vanish, T_1 is guaranteed not to vanish.

$$\begin{aligned} T_1 &= d(f_t)d(f_{t-1})\frac{\partial s_{t-2}}{\partial A} \\ &= d(f_t)d(f_{t-1})d(f_{t-2})\frac{\partial s_{t-3}}{\partial A} + \text{(other terms)} \\ &\vdots \\ &= \left(\prod_{j=k+1}^{t} d(f_j)\right)\frac{\partial s_k}{\partial A} + \text{(other terms)} \end{aligned} \tag{28}$$

In equation 28, we have unrolled the T_1 term to the k^{th} time-step. Let the 1^{st} term in the equation be called T_2. T_2 is one of the terms in $\frac{\partial h_t}{\partial A}$ that contributes

to the gradient for the state at a previous time-steps (in this case time-step k). We will now show that T_2 is sufficiently large, and does not vanish.
If the state at time-step k did not significantly contribute to the state at time-step t, then the term T_2 should indeed be close to zero. Note that this is unrelated to the vanishing gradient problem; the gradient should be legally zero if s_k does not contribute to s_t.
The interesting case is when s_k does contribute significantly to s_t. If this happens, then the values of the forget gate f_j should be close to $\mathbf{1}$ for all $j > k$ (if this does not happen, then at some time step the forget gate will be blocked, and s_k will not contribute to s_t, which leads to a contradiction). Thus, we will move forward with the assumption that $d(f_j) \approx \mathbf{1}$ for $j > k$. Let us look at T_2 with this in mind.

$$\begin{aligned} T_2 &= \left(\prod_{j=k+1}^{t} d(f_j) \right) \frac{\partial s_k}{\partial A} \\ &= D \frac{\partial s_k}{\partial A} \end{aligned} \tag{29}$$

Here D is a diagonal matrix defined by $D = d(f_{k+1} \odot f_{k+2} \ldots \odot f_t)$. The only way T_2 can vanish is if D formed by repeated multiplication vanishes. Since all f_j are close to $\mathbf{1}$, $\|D\| \approx \|I\| = \sqrt{n}$ (n is the LSTM state size). Therefore, T_2 does not vanish.

6 LSTMs do not solve the Exploding Gradient Problem

6.1 Problem Description

In RNNs, when the gradient of L_t w.r.t any parameter A is backpropagated through time, the norm of the gradient becomes exponentially larger, until it explodes to absurd values. This is a serious problem since the explosion forces us to truncate the propagation of gradients, preventing RNNs from learning long range temporal dependencies.

6.2 Intuitive Sketch of the Proof

If we are able to find any single path that explodes, then even if the other gradient flow paths don't explode, this particular path will cause the overall gradient to explode.
We will show the existence of such a common path in all the derivatives we have computed.
If we observe the derivative $\frac{\partial L_t}{\partial A}$ for any parameter A, then it can be written in the following form:

$$\frac{\partial h_t}{\partial A} = M_t \frac{h_{t-1}}{\partial A} + N_t \frac{\partial s_{t-1}}{\partial A} + V_t \tag{30}$$

Here M_t, N_t, V_t are tensors that are functions of the gates and states at time t. In the next section, we will show that the 1^{st} term $M_t \frac{h_{t-1}}{\partial A}$ could explode when it is unrolled to a large number of time steps.

6.3 Proof

Consider the 1^{st} term (represented as $M_t \frac{h_{t-1}}{\partial A}$ in equation 30) in each of the derivatives. Represent this term by T_0.

$$T_0 = \left(d(\sigma(s_t))d(o_t')W_o + d(o_t)d(\sigma'(s_t))(d(s_{t-1})d(f_t')W_f + d(i_t)d(\widetilde{s_t}')W + d(\widetilde{s}_t)d(i_t')W_i) \right) \frac{\partial h_{t-1}}{\partial A} \tag{31}$$

Consider the 1^{st} term of equation 31. Call this term T_1. We will show that this term by itself could explode.

$$\begin{aligned} T_1 &= d(\sigma(s_t))d(o_t')W_o \frac{\partial h_{t-1}}{\partial A} \\ &= d(\sigma(s_t))d(o_t')W_o d(\sigma(s_{t-1}))d(o_{t-1}')W_o \frac{\partial h_{t-2}}{\partial A} + (\text{other terms}) \\ &\vdots \\ &= \left(\prod_{j=k+1}^{t} \Big(d(\sigma(s_j))d(o_j')W_o \Big) \right) \frac{\partial h_k}{\partial A} + (\text{other terms}) \end{aligned} \tag{32}$$

In equation 32, we have unrolled the T_1 term to the k^{th} time-step. Let the 1^{st} term in the equation be called T_2. T_2 is one of the terms in $\frac{\partial h_t}{\partial A}$ that contributes to the gradient for the state at a previous time-steps (in this case time-step k). We will now show that T_2 could possibly explode.
Consider $\|d(\sigma(s_j))d(o_j')W_o\| \leq \|d(\sigma(s_j))\| \|d(o_j')\| \|W_o\| \leq \sqrt{n} * \frac{\sqrt{n}}{2} * \|W_o\| = \frac{n}{2}\|W_o\|$ (n is the LSTM state size).

$$\begin{aligned} \|T_2\| &= \| \left(\prod_{j=k+1}^{t} \Big(d(\sigma(s_j))d(o_j')W_o \Big) \right) \| \frac{\partial h_k}{\partial A} \| \\ &\leq \prod_{j=k+1}^{t} \Big(\frac{n}{2} \|W_o\| \Big) \| \frac{\partial h_k}{\partial A} \| \\ &\leq \Big(\frac{n}{2} \|W_o\| \Big)^{t-k} \| \frac{\partial h_k}{\partial A} \| \end{aligned} \tag{33}$$

The term $\left(\frac{n}{2}\|W_o\| \right)^{t-k}$ could explode if $\frac{n}{2}\|W_o\| > 1$, causing the overall gradient to explode.