LSTM: From theory to practice

Adrian Mihai Iosif, Matei Macri, Tudor Berariu*

January 2016

1 Introduction

This document is both a resource for understanding the mathematics of LSTM (Section 2) and a tutorial for step by step implementation in Torch (Section 3).

The LSTM module was first introduced in [1] as a solution to the *vanishing gradient* problem that made training vanilla RNNs difficult.

2 Mathematical foundations

Notations and other conventions

In this document the following notations are adopted:

	D	memory cell dimension
	N	input vector size
g_ι	$\mathbb{R}$	input gate
g_ϕ	$\mathbb{R}$	forget gate
g_ω	$\mathbb{R}$	output gate
$\mathbf{x}$	$\mathbb{R}^N$	inputs
$\mathbf{z}_c$	$\mathbb{R}^D$	input vector (a better name, maybe?)
$\mathbf{s}$	$\mathbb{R}^D$	the actual memory of the cell
$\mathbf{z}_h$	$\mathbb{R}^D$	output vector
$\mathbf{w}_\iota$	$\mathbb{R}^{(N+2D+1)}$	input gate parameters
$\mathbf{w}_\phi$	$\mathbb{R}^{(N+2D+1)}$	forget gate parameters
$\mathbf{w}_\omega$	$\mathbb{R}^{(N+2D+1)}$	output gate parameters
$\mathbf{W}_c$	$\mathbb{R}^{(N+D+1)\times D}$	input vector parameters
	f_*	activation functions (e.g. logistic) - applied element-wise

Vertical concatenation. We use a simplified notation for the vertical concatenation of two vectors $\mathbf{a}$, $\mathbf{b}$: $[\mathbf{a}; \mathbf{b}]$ instead of $\left[\mathbf{a}^{\mathrm{T}}; \mathbf{b}^{\mathrm{T}}\right]^{\mathrm{T}}$.

Element-wise multiplication. We use $\odot$ to denote element-wise multiplication of equally sized tensors.

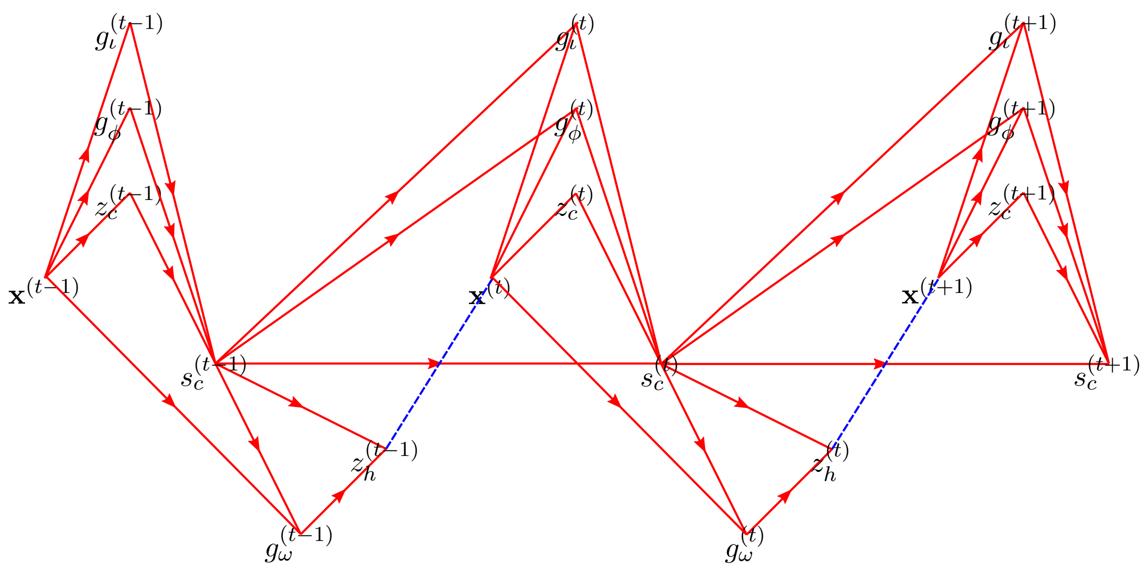
*This author wrote Section 2.

Unidimensional tensors. In this document we use the notation $\mathbf{v}$ for a column vector and $\mathbf{v}^{\mathrm{T}}$ whenever a row vector is needed. However, for various Jacobians we use a notation such as $\mathbf{j} = \frac{\partial \alpha}{\partial \mathbf{v}}$ for a row vector (a $1 \times size(\mathbf{v})$ Jacobian). Vectors that are not defined as Jacobians are always columns.

2.1 The Forward Phase

In what follows we describe the general formulas used to compute the output vector of the LSTM cell.

The computational graph



Input gate and input value

$$a_\iota^{(t)} = \underbrace{\mathbf{w}_\iota^{\mathrm{T}}}_{1\times(N+2D+1)} \cdot \underbrace{\left[\mathbf{x}^{(t)}; \mathbf{z}_h^{(t-1)}; \mathbf{s}^{(t-1)}; 1\right]}_{(N+2D+1)} \tag{1}$$

$$g_\iota^{(t)} = f_\iota\left(a_\iota^{(t)}\right) \tag{2}$$

$$\mathbf{a}_c^{(t)} = \underbrace{\mathbf{W}_c^{\mathrm{T}}}_{D\times(N+D+1)} \cdot \underbrace{\left[\mathbf{x}^{(t)}; \mathbf{z}_h^{(t-1)}; 1\right]}_{(N+D+1)} \tag{3}$$

$$\mathbf{z}_c^{(t)} = f_c\left(\mathbf{a}_c^{(t)}\right) \tag{4}$$

Forget gate

This gate actually acts as a *keep* gate.

$$a_\phi^{(t)} = \mathbf{w}_\phi^{\mathrm{T}} \cdot \left[\mathbf{x}^{(t)}; \mathbf{z}_h^{(t-1)}; \mathbf{s}^{(t-1)}; 1\right] \tag{5}$$

$$g_\phi^{(t)} = f_\phi\left(a_\phi^{(t)}\right) \tag{6}$$

Cell value

$$\mathbf{s}^{(t)} = g_\phi^{(t)}\mathbf{s}^{(t-1)} + g_\iota^{(t)}\mathbf{z}_c^{(t)} \tag{7}$$

Output gate and output value

$$a_\omega^{(t)} = \mathbf{w}_\omega^{\mathrm{T}} \cdot \left[\mathbf{x}^{(t)}; \mathbf{z}_h^{(t-1)}; \mathbf{s}^{(t)}; 1\right] \tag{8}$$

$$g_\omega^{(t)} = f_\omega\left(a_\omega^{(t)}\right) \tag{9}$$

$$\mathbf{z}_h^{(t)} = f_h\left(g_\omega^{(t)} f_s\left(\mathbf{s}^{(t)}\right)\right) \tag{10}$$

2.2 The Backward Phase

In this subsection we present the exact form of the partial derivatives of some error function with respect to the parameters of the LSTM cell.

Notations for various partial derivatives

The partial derivatives of the error E with respect to the parameters:

$$\boldsymbol{\delta}_\omega^{(t)} \stackrel{not.}{=} \frac{\partial E}{\partial \mathbf{w}_\omega} \qquad \boldsymbol{\delta}_\phi^{(t)} \stackrel{not.}{=} \frac{\partial E}{\partial \mathbf{w}_\phi} \qquad \boldsymbol{\delta}_\iota^{(t)} \stackrel{not.}{=} \frac{\partial E}{\partial \mathbf{w}_\iota} \tag{11}$$

$$\boldsymbol{\Delta}_c^{(t)} \stackrel{not.}{=} \frac{\partial E}{\partial \mathbf{W}_c} \tag{12}$$

The partial derivatives of the error E with respect to the gates:

$$\delta_{g_\omega}^{(t)} \stackrel{not.}{=} \frac{\partial E}{\partial g_\omega^{(t)}} \qquad \delta_{g_\phi}^{(t)} \stackrel{not.}{=} \frac{\partial E}{\partial g_\phi^{(t)}} \qquad \delta_{g_\iota}^{(t)} \stackrel{not.}{=} \frac{\partial E}{\partial g_\iota^{(t)}} \tag{13}$$

The partial derivatives of the error E with respect to $\mathbf{x}^{(t)}$ and with respect to $\mathbf{z}_h^{(t-1)}$:

$$\underbrace{\boldsymbol{\delta}_x^{(t)}}_{1\times N} \stackrel{not.}{=} \underbrace{\frac{\partial E}{\partial \mathbf{z}_h^{(t)}}}_{1\times D} \underbrace{\frac{\partial \mathbf{z}_h^{(t)}}{\partial \mathbf{x}^{(t)}}}_{D\times N} \tag{14}$$

$$\underbrace{\boldsymbol{\delta}_{zz}^{(t)\to(t-1)}}_{1\times D} \stackrel{not.}{=} \underbrace{\frac{\partial E}{\partial \mathbf{z}_h^{(t)}}}_{1\times D} \underbrace{\frac{\partial \mathbf{z}_h^{(t)}}{\partial \mathbf{z}_h^{(t-1)}}}_{D\times D} \tag{15}$$

$$\boldsymbol{\delta}_s^{(t)} \stackrel{not.}{=} \frac{\partial E}{\partial \mathbf{s}^{(t)}} \qquad \boldsymbol{\delta}_{\mathbf{z}_c}^{(t)} \stackrel{not.}{=} \frac{\partial E}{\partial \mathbf{z}_c^{(t)}} \qquad \boldsymbol{\delta}_{\mathbf{z}_h}^{(t)} \stackrel{not.}{=} \frac{\partial E}{\partial \mathbf{z}_h^{(t)}} \tag{16}$$

Inner loops

Before computing the needed gradients, let's take a closer look at $\frac{\partial E}{\partial \mathbf{s}^{(t)}}$. This gradient has two components. The first corresponds to the error flowing through $\mathbf{z}_h^{(t)}$ and the second corresponds to the inner loops of the LSTM (the connections to $g_\iota^{(t+1)}$, $g_\phi^{(t+1)}$, and $\mathbf{s}^{(t+1)}$).

$$
\begin{aligned}
\boldsymbol{\delta}_s^{(t)} \overset{not.}{=} \frac{\partial E}{\partial \mathbf{s}^{(t)}} =& \frac{\partial E}{\partial \mathbf{z}_h^{(t)}} \frac{\partial \mathbf{z}_h^{(t)}}{\partial \mathbf{s}^{(t)}} + \frac{\partial E}{\partial g_\phi^{(t+1)}} \frac{\partial g_\phi^{(t+1)}}{\partial \mathbf{s}^{(t)}} + \frac{\partial E}{\partial g_\iota^{(t+1)}} \frac{\partial g_\iota^{(t+1)}}{\partial \mathbf{s}^{(t)}} + \frac{\partial E}{\partial \mathbf{s}^{(t+1)}} \frac{\partial \mathbf{s}^{(t+1)}}{\partial \mathbf{s}^{(t)}} \\
=& \frac{\partial E}{\partial \mathbf{z}_h^{(t)}} \frac{\partial \mathbf{z}_h^{(t)}}{\partial \left(g_w^{(t)} f_s \left(\mathbf{s}^{(t)} \right) \right)} \frac{\partial \left(g_w^{(t)} \mathbf{s}^{(t)} \right)}{\partial \mathbf{s}^{(t)}} \\
&+ \frac{\partial E}{\partial g_\phi^{(t+1)}} \frac{\partial g_\phi^{(t+1)}}{\partial \mathbf{s}^{(t)}} + \frac{\partial E}{\partial g_\iota^{(t+1)}} \frac{\partial g_\iota^{(t+1)}}{\partial \mathbf{s}^{(t)}} + \frac{\partial E}{\partial \mathbf{s}^{(t+1)}} \frac{\partial \mathbf{s}^{(t+1)}}{\partial \mathbf{s}^{(t)}} \\
=& \underbrace{\boldsymbol{\delta}_{\mathbf{z}_h}^{(t)}}_{1\times D} \underbrace{\operatorname{diag}\left(f_h' \left(g_\omega^{(t)} f_s \left(\mathbf{s}^{(t)} \right) \right) \right)}_{D\times D} \underbrace{\left(g_\omega^{(t)} \operatorname{diag}\left(f_s' \left(\mathbf{s}^{(t)} \right) \right) + f_\omega' \left(a_\omega \right) f_s \left(\mathbf{s}^{(t)} \right) \mathbf{w}_{\omega,s}^{\mathrm{T}} \right)}_{D\times D} + \\
&+ \delta_{g_\phi}^{(t+1)} f_\phi' \left(a_\phi^{(t+1)} \right) \mathbf{w}_{\phi,s}^{\mathrm{T}} + \delta_{g_\iota}^{(t+1)} f_\iota' \left(a_\iota^{(t+1)} \right) \mathbf{w}_{\iota,s}^{\mathrm{T}} + g_\phi^{(t+1)} \boldsymbol{\delta}_s^{(t+1)\mathrm{T}} \\
=& \boldsymbol{\delta}_s^{(t)\to(t)} + \boldsymbol{\delta}_s^{(t+1)\to(t)}
\end{aligned}
\tag{17}
$$

It's easier to understand the above matriceal expression by observing an element of the $1 \times D$ row vector $\frac{\partial E}{\partial \mathbf{z}_h^{(t)}} \frac{\partial \mathbf{z}_h^{(t)}}{\partial \mathbf{s}^{(t)}}$:

$$
\begin{aligned}
\boldsymbol{\delta}_s^{(t)\to(t)}[i] &= \sum_{j=1} \frac{\partial E}{\partial \mathbf{z}_h^{(t)}[j]} \frac{\partial \mathbf{z}_h^{(t)}[j]}{\partial \mathbf{s}^{(t)}[i]} \\
&= \sum_{j=1} \left[\frac{\partial E}{\partial \mathbf{z}_h^{(t)}[j]} f_h' \left(g_\omega^{(t)} f_s \left(\mathbf{s}^{(t)}[j] \right) \right) \left(g_\omega^{(t)} \delta_{i,j} f_s' \left(\mathbf{s}^{(t)}[i] \right) + f_s \left(\mathbf{s}^{(t)}[j] \right) f_\omega' \left(a_\omega^{(t)} \right) \mathbf{w}_{\omega,s}[i] \right) \right]
\end{aligned}
\tag{18}
$$

In the above expression $\delta_{i,j}$ is the Kronecker-Delta function.

The objective of the backward phase

Given $\boldsymbol{\delta}_{\mathbf{z}_h}^{(t)}$, and $\boldsymbol{\delta}_s^{(t+1)\to(t)}$, the following derivatives need to be computed: $\boldsymbol{\delta}_\omega^{(t)}$, $\boldsymbol{\delta}_\phi^{(t)}$, $\boldsymbol{\delta}_\iota^{(t)}$, $\boldsymbol{\Delta}_c^{(t)}$ (for optimization purposes) $\boldsymbol{\delta}_x^{(t)}$, $\boldsymbol{\delta}_{zz}^{(t)\to(t-1)}$, and $\boldsymbol{\delta}_s^{(t)\to(t-1)}$ (in order to compute other gradients back in the architecture).

Order of computation

Gradients need to be computed in the following order: $\delta_{g_\omega}^{(t)}$, $\boldsymbol{\delta}_\omega^{(t)}$, $\boldsymbol{\delta}_{\mathbf{s}}^{(t)\to(t)}$, $\boldsymbol{\delta}_{\mathbf{s}}^{(t)}$, $\boldsymbol{\delta}_{\mathbf{z}_c}^{(t)}$, $\boldsymbol{\Delta}_c^{(t)}$, δ_{g_ι}, $\boldsymbol{\delta}_\iota^{(t)}$, δ_{g_ϕ}, $\boldsymbol{\delta}_\phi^{(t)}$, $\boldsymbol{\delta}_s^{(t)\to(t-1)}$, $\boldsymbol{\delta}_x^{(t)}$, $\boldsymbol{\delta}_{zz}^{(t)\to(t-1)}$.

The gradients with respect to the output gate and its parameters

$$\delta_{g_\omega}^{(t)} \overset{\text{not.}}{=} \frac{\partial E}{\partial g_\omega^{(t)}} = \frac{\partial E}{\partial \mathbf{z}_h^{(t)}} \frac{\partial \mathbf{z}_h^{(t)}}{\partial g_\omega^{(t)}} = \boldsymbol{\delta}_{\mathbf{z}_h}^{(t)} \left(f_h' \left(g_\omega^{(t)} f_s \left(\mathbf{s}^{(t)} \right) \right) \odot \mathbf{s}^{(t)} \right) \tag{19}$$

$$\boldsymbol{\delta}_\omega^{(t)} \overset{\text{not.}}{=} \frac{\partial E}{\partial \mathbf{w}_\omega^{(t)}} = \frac{\partial E}{\partial g_\omega^{(t)}} \frac{\partial g_\omega^{(t)}}{\partial \mathbf{w}_\omega^{(t)}} = \delta_{g_\omega}^{(t)} f_\omega' \left(a_\omega^{(t)} \right) \left[\mathbf{x}^{(t)}; \mathbf{z}_h^{(t-1)}; \mathbf{s}^{(t)}; 1 \right] \tag{20}$$

The gradients with respect to the memory cell

$$\begin{aligned}
\boldsymbol{\delta}_s^{(t)\to(t)} &\overset{\text{not.}}{=} \frac{\partial E}{\partial \mathbf{z}_h^{(t)}} \frac{\partial \mathbf{z}_h^{(t)}}{\mathbf{s}^{(t)}} &(21)\\
&= \boldsymbol{\delta}_{\mathbf{z}_h}^{(t)} \operatorname{diag}\left(f_h' \left(g_\omega^{(t)} f_s \left(\mathbf{s}^{(t)} \right) \right) \right) \left(g_\omega^{(t)} \operatorname{diag}\left(f_s' \left(\mathbf{s}^{(t)} \right) \right) + f_\omega' \left(a_\omega \right) f_s \left(\mathbf{s}^{(t)} \right) \mathbf{w}_{\omega,s}^{\mathrm{T}} \right) &(22)\\
&= \left(\boldsymbol{\delta}_{\mathbf{z}_h}^{(t)} \odot f_h' \left(g_\omega^{(t)} f_s \left(\mathbf{s}^{(t)} \right) \right)^{\mathrm{T}} \right) \left(g_\omega^{(t)} \operatorname{diag}\left(f_s' \left(\mathbf{s}^{(t)} \right) \right) + f_\omega' \left(a_\omega \right) f_s \left(\mathbf{s}^{(t)} \right) \mathbf{w}_{\omega,s}^{\mathrm{T}} \right) &(23)\\
\boldsymbol{\delta}_s^{(t)} &\overset{\text{not.}}{=} \boldsymbol{\delta}_s^{(t)\to(t)} + \boldsymbol{\delta}_s^{(t+1)\to(t)} &(24)
\end{aligned}$$

The gradients with respect to the input value and its parameters

$$\boldsymbol{\delta}_{\mathbf{z}_c}^{(t)} \overset{\text{not.}}{=} \frac{\partial E}{\partial \mathbf{z}_c^{(t)}} = \frac{\partial E}{\partial \mathbf{s}^{(t)}} \frac{\partial \mathbf{s}^{(t)}}{\partial \mathbf{z}_c^{(t)}} = \boldsymbol{\delta}_s^{(t)} \left(g_i^{(t)} \mathbf{I}_D \right) = g_i^{(t)} \boldsymbol{\delta}_s^{(t)} \tag{25}$$

$$\boldsymbol{\Delta}_c^{(t)} \overset{\text{not.}}{=} \frac{\partial E}{\partial \mathbf{W}_c^{(t)}} = \frac{\partial E}{\partial \mathbf{z}_c^{(t)}} \frac{\partial \mathbf{z}_c^{(t)}}{\partial \mathbf{a}_c^{(t)}} \frac{\partial \mathbf{a}_c^{(t)}}{\partial \mathbf{W}_c^{(t)}} = \underbrace{\left({\boldsymbol{\delta}_{\mathbf{z}_c}^{(t)}}^{\mathrm{T}} \odot f_c' \left(\mathbf{a}_c \right) \right)}_{D\times 1} \underbrace{\left[{\mathbf{x}^{(t)}}^{\mathrm{T}}; {\mathbf{s}^{(t-1)}}^{\mathrm{T}}; 1 \right]}_{1\times(N+D+1)} \tag{26}$$

The gradients with respect to the input gate and its parameters

$$\delta_{g_\iota}^{(t)} \overset{\text{not.}}{=} \frac{\partial E}{\partial g_\iota^{(t)}} = \frac{\partial E}{\partial \mathbf{s}^{(t)}} \frac{\partial \mathbf{s}^{(t)}}{\partial g_\iota^{(t)}} = \boldsymbol{\delta}_s^{(t)} \mathbf{z}_c^{(t)} \tag{27}$$

$$\boldsymbol{\delta}_\iota^{(t)} \overset{\text{not.}}{=} \frac{\partial E}{\partial \mathbf{w}_\iota^{(t)}} = \frac{\partial E}{\partial g_\iota^{(t)}} \frac{\partial g_\iota^{(t)}}{\partial \mathbf{w}_\iota^{(t)}} = \delta_{g_\iota}^{(t)} f_\iota' \left(a_\iota^{(t)} \right) \left[{\mathbf{x}^{(t)}}^{\mathrm{T}}; {\mathbf{z}_h^{(t-1)}}^{\mathrm{T}}; {\mathbf{s}^{(t-1)}}^{\mathrm{T}}; 1 \right] \tag{28}$$

The gradients with respect to the forget gate and its parameters

$$\delta_{g_\phi}^{(t)} \overset{\text{not.}}{=} \frac{\partial E}{\partial g_\phi^{(t)}} = \frac{\partial E}{\partial \mathbf{s}^{(t)}} \frac{\partial \mathbf{s}^{(t)}}{\partial g_\phi^{(t)}} = \boldsymbol{\delta}_s^{(t)} \mathbf{s}^{(t-1)} \tag{29}$$

$$\boldsymbol{\delta}_\phi^{(t)} \overset{\text{not.}}{=} \frac{\partial E}{\partial \mathbf{w}_\phi^{(t)}} = \frac{\partial E}{\partial g_\phi^{(t)}} \frac{\partial g_\phi^{(t)}}{\partial \mathbf{w}_\phi^{(t)}} = \delta_{g_\phi}^{(t)} f_\phi' \left(a_\phi^{(t)} \right) \left[{\mathbf{x}^{(t)}}^{\mathrm{T}}; {\mathbf{z}_h^{(t-1)}}^{\mathrm{T}}; {\mathbf{s}^{(t-1)}}^{\mathrm{T}}; 1 \right] \tag{30}$$

The gradients with respect to incoming values

$$\boldsymbol{\delta}_s^{(t)\to(t-1)} = \delta_{g_\phi}^{(t)} f'_\phi \left(a_\phi^{(t)}\right) \mathbf{w}_{\phi,s}^{\mathrm{T}} + \delta_{g_\iota}^{(t)} f'_\iota \left(a_\iota^{(t)}\right) \mathbf{w}_{\iota,s}^{\mathrm{T}} + g_\phi^{(t)} {\boldsymbol{\delta}_s^{(t)}}^{\mathrm{T}} \tag{31}$$

$$\begin{aligned}
\boldsymbol{\delta}_x^{(t)} &\overset{\text{not.}}{=} \frac{\partial E}{\partial \mathbf{z}_h^{(t)}} \frac{\partial \mathbf{z}_h^{(t)}}{\partial \mathbf{x}^{(t)}} \\
&= \frac{\partial E}{\partial \mathbf{z}_h^{(t)}} \left(\frac{\partial \mathbf{z}_h^{(t)}}{\partial g_\iota^{(t)}} \frac{\partial g_\iota^{(t)}}{\partial \mathbf{x}^{(t)}} + \frac{\partial \mathbf{z}_h^{(t)}}{\partial g_\phi^{(t)}} \frac{\partial g_\phi^{(t)}}{\partial \mathbf{x}^{(t)}} + \frac{\partial \mathbf{z}_h^{(t)}}{\partial g_\omega^{(t)}} \frac{\partial g_\omega^{(t)}}{\partial \mathbf{x}^{(t)}} + \frac{\partial \mathbf{z}_h^{(t)}}{\partial \mathbf{z}_c^{(t)}} \frac{\partial \mathbf{z}_c^{(t)}}{\partial \mathbf{x}^{(t)}} \right) \\
&= \delta_{g_\iota}^{(t)} f'_\iota \left(a_\iota^{(t)}\right) \mathbf{w}_{\iota,x}^{\mathrm{T}} + \delta_{g_\phi}^{(t)} f'_\phi \left(a_\phi^{(t)}\right) \mathbf{w}_{\phi,x}^{\mathrm{T}} + \delta_{g_\omega}^{(t)} f'_\omega \left(a_\omega^{(t)}\right) \mathbf{w}_{\omega,x}^{\mathrm{T}} + \underbrace{\left(\boldsymbol{\delta}_{\mathbf{z}_c}^{(t)} \odot f'_c \left(\mathbf{a}_c^{(t)}\right)^{\mathrm{T}}\right)}_{1\times D} \underbrace{\mathbf{W}_{c,x}^{\mathrm{T}}}_{D\times N}
\end{aligned} \tag{32}$$

$$\begin{aligned}
\boldsymbol{\delta}_{zz}^{(t)\to(t-1)} &\overset{\text{not.}}{=} \frac{\partial E}{\partial \mathbf{z}_h^{(t)}} \frac{\partial \mathbf{z}_h^{(t)}}{\partial \mathbf{z}_h^{(t-1)}} \\
&= \frac{\partial E}{\partial \mathbf{z}_h^{(t)}} \left(\frac{\partial \mathbf{z}_h^{(t)}}{\partial g_\iota^{(t)}} \frac{\partial g_\iota^{(t)}}{\partial \mathbf{z}_h^{(t-1)}} + \frac{\partial \mathbf{z}_h^{(t)}}{\partial g_\phi^{(t)}} \frac{\partial g_\phi^{(t)}}{\partial \mathbf{z}_h^{(t-1)}} + \frac{\partial \mathbf{z}_h^{(t)}}{\partial g_\omega^{(t)}} \frac{\partial g_\omega^{(t)}}{\partial \mathbf{z}_h^{(t-1)}} + \frac{\partial \mathbf{z}_h^{(t)}}{\partial \mathbf{z}_c^{(t)}} \frac{\partial \mathbf{z}_c^{(t)}}{\partial \mathbf{z}_h^{(t-1)}} \right) \\
&= \delta_{g_\iota}^{(t)} f'_\iota \left(a_\iota^{(t)}\right) \mathbf{w}_{\iota,z}^{\mathrm{T}} + \delta_{g_\phi}^{(t)} f'_\phi \left(a_\phi^{(t)}\right) \mathbf{w}_{\phi,z}^{\mathrm{T}} + \delta_{g_\omega}^{(t)} f'_\omega \left(a_\omega^{(t)}\right) \mathbf{w}_{\omega,z}^{\mathrm{T}} + \\
&\quad + \underbrace{\left(\boldsymbol{\delta}_{\mathbf{z}_c}^{(t)} \odot f'_c \left(\mathbf{a}_c^{(t)}\right)^{\mathrm{T}}\right)}_{1\times D} \underbrace{\mathbf{W}_{c,z}^{\mathrm{T}}}_{D\times D}
\end{aligned} \tag{33}$$

3 Torch Implementation

TODO

References

[1] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory”. In: *Neural computation* 9.8 (1997), pp. 1735–1780.