LSTM: From theory to practice

Adrian Mihai Iosif, Matei Macri, Tudor Berariu

January 2016

1 Introduction

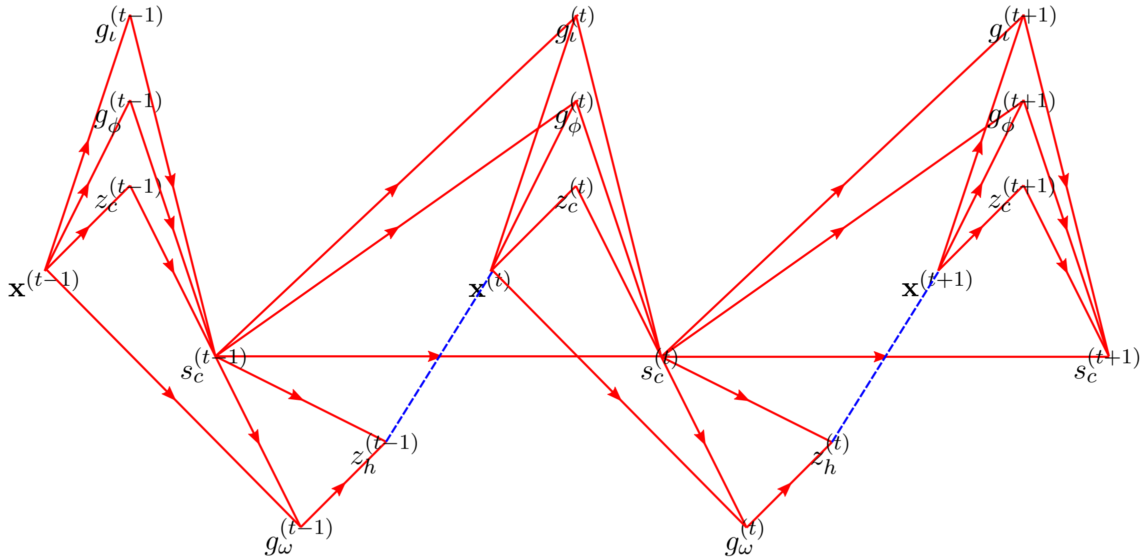
This document is a both a resource for understanding the mathematics of LSTM (Section 2) and a tutorial containing a step by step implementation in Torch (Section 3). The LSTM module was first introduced in [1]. It solves the *vanishing gradient* problem that makes training vanilla RNNs difficult.

2 Mathematical foundations

Conventions. We propose the following notation for the variables:

- g_* - gates
- z_c / z_h - input / output values
- $\mathbf{w}_*$ - weights (parameters)
- f_* - activation functions (e.g. logistic)
- $\mathbf{x}$ - inputs
- s - a scalar value (the actual memory of the LSTM cell)

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



2.1 The Forward Phase

Input gate and input value.

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

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

$$a_c^{(t)} = \mathbf{w}_c^{\mathrm{T}} \cdot \left[\mathbf{x}^{(t)}; z_h^{(t-1)}; 1\right] \tag{3}$$

$$z_c^{(t)} = f_c\left(a_c^{(t)}\right) \tag{4}$$

Forget gate. This is actually a *keep* gate:

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

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

Cell value.

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

Output gate and output value.

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

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

$$z_h^{(t)} = f_h\left(g_\omega^{(t)} s^{(t)}\right) \tag{10}$$

2.2 The Backward Phase

Notations. In what follows the following notations are used for various partial derivatives:

$$\boldsymbol{\delta}_x^{(t)} \overset{not.}{=} \frac{\partial E}{\partial z_h^{(t)}} \frac{\partial z_h^{(t)}}{\partial \mathbf{x}^{(t)}}; \quad \delta_{zz}^{(t)\to(t-1)} \overset{not.}{=} \frac{\partial E}{\partial z_h^{(t)}} \frac{\partial z_h^{(t)}}{\partial z_h^{(t-1)}};$$

$$\delta_s^{(t)} \overset{not.}{=} \frac{\partial E}{\partial s^{(t)}}; \quad \delta_c^{(t)} \overset{not.}{=} \frac{\partial E}{\partial z_c^{(t)}}; \quad \delta_h^{(t)} \overset{not.}{=} \frac{\partial E}{\partial z_h^{(t)}};$$

$$\delta_{g_\omega}^{(t)} \overset{not.}{=} \frac{\partial E}{\partial g_\omega^{(t)}}; \quad \delta_{g_\phi}^{(t)} \overset{not.}{=} \frac{\partial E}{\partial g_\phi^{(t)}}; \quad \delta_{g_\iota}^{(t)} \overset{not.}{=} \frac{\partial E}{\partial g_\iota^{(t)}}$$

$$\boldsymbol{\delta}_\omega^{(t)} \overset{not.}{=} \frac{\partial E}{\partial \mathbf{w}_\omega}; \quad \boldsymbol{\delta}_\phi^{(t)} \overset{not.}{=} \frac{\partial E}{\partial \mathbf{w}_\phi}; \quad \boldsymbol{\delta}_\iota^{(t)} \overset{not.}{=} \frac{\partial E}{\partial \mathbf{w}_\iota}$$

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

$$\begin{aligned}
\delta_s^{(t)} \overset{not.}{=} \frac{\partial E}{\partial s^{(t)}} &= \frac{\partial E}{\partial z_h^{(t)}} \frac{\partial z_h^{(t)}}{\partial s^{(t)}} + \frac{\partial E}{\partial g_\phi^{(t+1)}} \frac{\partial g_\phi^{(t+1)}}{\partial s^{(t)}} + \frac{\partial E}{\partial g_\iota^{(t+1)}} \frac{\partial g_\iota^{(t+1)}}{\partial s^{(t)}} + \frac{\partial E}{\partial s^{(t+1)}} \frac{\partial s^{(t+1)}}{\partial s^{(t)}} \\
&= \delta_h^{(t)} f'_\omega \left(g_\omega^{(t)} s^{(t)}\right) g_\omega^{(t)} + \delta_{g_\phi}^{(t+1)} f'_\phi \left(a_\phi^{(t+1)}\right) w_{\phi,s} + \delta_{g_\iota}^{(t+1)} f'_\iota \left(a_\iota^{(t+1)}\right) w_{\iota,s} + \delta_s^{(t+1)} g_\phi^{(t+1)} \\
&= \delta_s^{(t)\to(t)} + \delta_s^{(t+1)\to(t)}
\end{aligned} \tag{11}$$

Goal. Given $\delta_h^{(t)}$, and $\delta_s^{(t+1)\to(t)}$, the following derivatives need to be computed: $\dfrac{\partial E}{\partial W_*}$, $\boldsymbol{\delta}_x^{(t)}$, $\delta_{zz}^{(t)\to(t-1)}$, and $\delta_s^{(t)\to(t-1)}$.

Order of computation. Gradients need to be computed in the following order: $\delta_{g_\omega}^{(t)}$, $\boldsymbol{\delta}_\omega^{(t)}$, $\delta_s^{(t)\to(t)}$, $\delta_s^{(t)}$, $\delta_c^{(t)}$, δ_{g_ι}, $\boldsymbol{\delta}_\iota^{(t)}$, δ_{g_ϕ}, $\boldsymbol{\delta}_\phi^{(t)}$, $\delta_s^{(t)\to(t-1)}$, $\boldsymbol{\delta}_x^{(t)}$, $\delta_{zz}^{(t)\to(t-1)}$.

$$\delta_{g_\omega}^{(t)} \overset{\text{not.}}{=} \frac{\partial E}{\partial g_\omega^{(t)}} = \frac{\partial E}{\partial z_h^{(t)}} \frac{\partial z_h^{(t)}}{\partial g_\omega^{(t)}} = \delta_h^{(t)} f_h' \left(g_\omega^{(t)} s^{(t)} \right) s^{(t)} \tag{12}$$

$$\boldsymbol{\delta}_\omega^{(t)} \overset{\text{not.}}{=} \frac{\partial E}{\partial \boldsymbol{w}_\omega^{(t)}} = \frac{\partial E}{\partial g_\omega^{(t)}} \frac{\partial g_\omega^{(t)}}{\partial \boldsymbol{w}_\omega^{(t)}} = \delta_{g_\omega}^{(t)} f_\omega' \left(a_\omega^{(t)} \right) \cdot \left[\mathbf{x}^{(t)}; z_h^{(t-1)}; s^{(t)}; 1 \right] \tag{13}$$

$$\delta_s^{(t)\to(t)} \overset{\text{not.}}{=} \frac{\partial E}{\partial z_h^{(t)}} \frac{\partial z_h^{(t)}}{s^{(t)}} = \delta_h^{(t)} f_\omega' \left(g_\omega^{(t)} s^{(t)} \right) g_\omega^{(t)} \tag{14}$$

$$\delta_s^{(t)} \overset{\text{not.}}{=} \delta_s^{(t)\to(t)} + \delta_s^{(t+1)\to(t)} \tag{15}$$

$$\delta_c^{(t)} \overset{\text{not.}}{=} \frac{\partial E}{\partial z_c^{(t)}} = \frac{\partial E}{\partial s^{(t)}} \frac{\partial s^{(t)}}{z_c^{(t)}} = \delta_s^{(t)} g_i^{(t)} \tag{16}$$

$$\delta_{g_\iota}^{(t)} \overset{\text{not.}}{=} \frac{\partial E}{\partial g_\iota^{(t)}} = \frac{\partial E}{\partial s^{(t)}} \frac{\partial s^{(t)}}{g_\iota^{(t)}} = \delta_s^{(t)} z_c^{(t)} \tag{17}$$

$$\boldsymbol{\delta}_\iota^{(t)} \overset{\text{not.}}{=} \frac{\partial E}{\partial \boldsymbol{w}_\iota^{(t)}} = \frac{\partial E}{\partial g_\iota^{(t)}} \frac{\partial g_\iota^{(t)}}{\partial \boldsymbol{w}_\iota^{(t)}} = \delta_{g_\phi}^{(t)} f_\iota' \left(a_\iota^{(t)} \right) \cdot \left[\mathbf{x}^{(t)}; z_h^{(t-1)}; s^{(t-1)}; 1 \right] \tag{18}$$

$$\delta_{g_\phi}^{(t)} \overset{\text{not.}}{=} \frac{\partial E}{\partial g_\phi^{(t)}} = \frac{\partial E}{\partial s^{(t)}} \frac{\partial s^{(t)}}{g_\phi^{(t)}} = \delta_s^{(t)} s^{(t-1)} \tag{19}$$

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

$$\delta_s^{(t)\to(t-1)} = \delta_{g_\phi}^{(t)} f_\phi' \left(a_\phi^{(t)} \right) w_{\phi,s} + \delta_{g_\iota}^{(t)} f_\iota' \left(a_\iota^{(t)} \right) w_{\iota,s} + \delta_s^{(t)} g_\phi^{(t)} \tag{21}$$

$$\begin{aligned} \boldsymbol{\delta}_x^{(t)} &\overset{\text{not.}}{=} \frac{\partial E}{\partial z_h^{(t)}} \frac{\partial z_h^{(t)}}{\partial \mathbf{x}^{(t)}} \\ &= \frac{\partial E}{\partial z_h^{(t)}} \left(\frac{\partial z_h^{(t)}}{\partial g_\iota^{(t)}} \frac{\partial g_\iota^{(t)}}{\partial \mathbf{x}^{(t)}} + \frac{\partial z_h^{(t)}}{\partial g_\phi^{(t)}} \frac{\partial g_\phi^{(t)}}{\partial \mathbf{x}^{(t)}} + \frac{\partial z_h^{(t)}}{\partial g_\omega^{(t)}} \frac{\partial g_\omega^{(t)}}{\partial \mathbf{x}^{(t)}} + \frac{\partial z_h^{(t)}}{\partial z_c^{(t)}} \frac{\partial z_c^{(t)}}{\partial \mathbf{x}^{(t)}} \right) \\ &= \delta_{g_\iota}^{(t)} f_\iota' \left(a_\iota^{(t)} \right) \mathbf{w}_{\iota,x} + \delta_{g_\phi}^{(t)} f_\phi' \left(a_\phi^{(t)} \right) \mathbf{w}_{\phi,x} + \delta_{g_\omega}^{(t)} f_\omega' \left(a_\omega^{(t)} \right) \mathbf{w}_{\omega,x} + \delta_c^{(t)} f_c' \left(a_c^{(t)} \right) \mathbf{w}_{c,x} \end{aligned} \tag{22}$$

$$\begin{aligned} \boldsymbol{\delta}_x^{(t)} &\overset{\text{not.}}{=} \frac{\partial E}{\partial z_h^{(t)}} \frac{\partial z_h^{(t)}}{\partial z_h^{(t-1)}} \\ &= \frac{\partial E}{\partial z_h^{(t)}} \left(\frac{\partial z_h^{(t)}}{\partial g_\iota^{(t)}} \frac{\partial g_\iota^{(t)}}{\partial z_h^{(t-1)}} + \frac{\partial z_h^{(t)}}{\partial g_\phi^{(t)}} \frac{\partial g_\phi^{(t)}}{\partial z_h^{(t-1)}} + \frac{\partial z_h^{(t)}}{\partial g_\omega^{(t)}} \frac{\partial g_\omega^{(t)}}{\partial z_h^{(t-1)}} + \frac{\partial z_h^{(t)}}{\partial z_c^{(t)}} \frac{\partial z_c^{(t)}}{\partial z_h^{(t-1)}} \right) \\ &= \delta_{g_\iota}^{(t)} f_\iota' \left(a_\iota^{(t)} \right) \mathbf{w}_{\iota,z} + \delta_{g_\phi}^{(t)} f_\phi' \left(a_\phi^{(t)} \right) \mathbf{w}_{\phi,z} + \delta_{g_\omega}^{(t)} f_\omega' \left(a_\omega^{(t)} \right) \mathbf{w}_{\omega,z} + \delta_c^{(t)} f_c' \left(a_c^{(t)} \right) \mathbf{w}_{c,z} \end{aligned} \tag{23}$$

3 Torch Implementation

References

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