Long-short Term Memory (LSTM): back-propagation derivation

Harsha Vardhan

May 4, 2022

Abstract

This document contains derivation of the gradients for a Long-short Term Memory Unit, using back-propagation (or reverse-mode differentiation). For the implementation of the neural-network see the accompanying notebooks.

1 Block Architecture

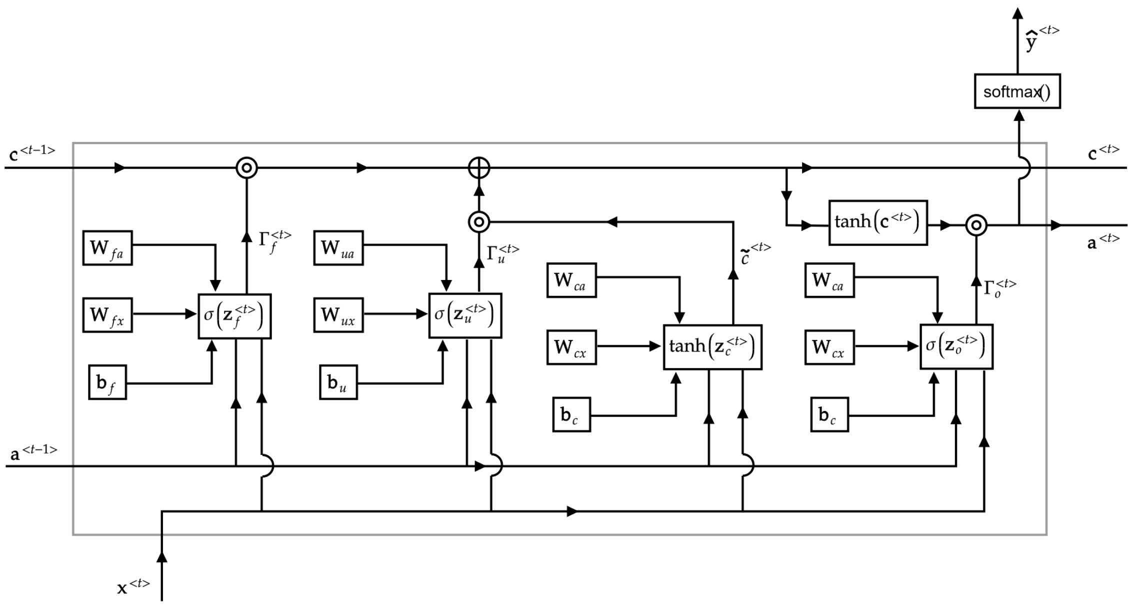

Figure 1: forward propagation diagram for a LSTM-block at time-step t. The concentric circles represent a hadamard-product (i.e. $\mathbf{x} \circ \mathbf{y}$) of the input vectors.

2 Forward Propagation

Given, $\mathbf{c}^{\langle t-1\rangle}$ and $\mathbf{a}^{\langle t-1\rangle}$ from the $(t-1)^{th}$ time-step, the equations for forward propagation through the t^{th} time-step, are as follows: (see figure-1)

$$
\begin{aligned}
\mathbf{z}_f^{\langle t \rangle} &= \mathbf{W}_{fa}^{\intercal}\,\mathbf{a}^{\langle t-1 \rangle} + \mathbf{W}_{fx}^{\intercal}\,\mathbf{x}^{\langle t \rangle} + \mathbf{b}_f && \text{eq.1} \\
\Gamma_f^{\langle t \rangle} &= \sigma(\mathbf{z}_f^{\langle t \rangle}) && \text{eq.2} \\
\mathbf{z}_u^{\langle t \rangle} &= \mathbf{W}_{ua}^{\intercal}\,\mathbf{a}^{\langle t-1 \rangle} + \mathbf{W}_{ux}^{\intercal}\,\mathbf{x}^{\langle t \rangle} + \mathbf{b}_u && \text{eq.3} \\
\Gamma_u^{\langle t \rangle} &= \sigma(\mathbf{z}_u^{\langle t \rangle}) && \text{eq.4} \\
\mathbf{z}_o^{\langle t \rangle} &= \mathbf{W}_{oa}^{\intercal}\,\mathbf{a}^{\langle t-1 \rangle} + \mathbf{W}_{ox}^{\intercal}\,\mathbf{x}^{\langle t \rangle} + \mathbf{b}_o && \text{eq.5} \\
\Gamma_o^{\langle t \rangle} &= \sigma(\mathbf{z}_o^{\langle t \rangle}) && \text{eq.6} \\
\mathbf{z}_c^{\langle t \rangle} &= \mathbf{W}_{ca}^{\intercal}\,\mathbf{a}^{\langle t-1 \rangle} + \mathbf{W}_{cx}^{\intercal}\,\mathbf{x}^{\langle t \rangle} + \mathbf{b}_c && \text{eq.7} \\
\tilde{\mathbf{c}}^{\langle t \rangle} &= \tanh(\mathbf{z}_c^{\langle t \rangle}) && \text{eq.8} \\
\mathbf{c}^{\langle t \rangle} &= \Gamma_u \circ \tilde{\mathbf{c}}^{\langle t \rangle} + \Gamma_f \circ \mathbf{c}^{\langle t-1 \rangle} && \text{eq.9} \\
\mathbf{a}^{\langle t \rangle} &= \tanh(\mathbf{c}^{\langle t \rangle}) \circ \Gamma_o && \text{eq.10} \\
\hat{\mathbf{y}}^{\langle t \rangle} &= \mathrm{softmax}(\mathbf{a}^{\langle t \rangle}) && \text{eq.11}
\end{aligned}
$$

where,

$$
\begin{aligned}
\mathbf{x}^{\langle t \rangle} &\text{ is a } (n_x, 1)\text{-dimensional vector} \\
\mathbf{c}^{\langle t \rangle} \,\&\, \mathbf{c}^{\langle t-1 \rangle} \,\&\, \tilde{\mathbf{c}}^{\langle t \rangle} &\text{ are } (n_a, 1)\text{-dimensional vectors} \\
\mathbf{a}^{\langle t \rangle} \,\&\, \mathbf{a}^{\langle t-1 \rangle} &\text{ are } (n_a, 1)\text{-dimensional vectors} \\
\mathbf{W}_{*a}, * \in \{f, u, o, c\} &\text{ are } (n_a, n_a)\text{-dimensional parameter matrices} \\
\mathbf{W}_{*x}, * \in \{f, u, o, c\} &\text{ are } (n_x, n_a)\text{-dimensional parameter matrices} \\
\mathbf{b}_{*}, * \in \{f, u, o, c\} &\text{ are } (n_a, 1)\text{-dimensional bias-vectors} \\
\Gamma_*^{\langle t \rangle}, * \in \{f, u, o\} &\text{ are } (n_a, 1)\text{-dimensional vector, which represents a gate} \\
\mathbf{z}_*^{\langle t \rangle}, * \in \{f, u, o, c\} &\text{ are } (n_a, 1)\text{-dimensional vector, is input to a gates' activation}
\end{aligned}
$$

Also, the sub-scripts denote the following: $_o$ - output-gate, $_u$ - update-gate, and $_f$ - forget-gate.

2.1 Backward Propagation

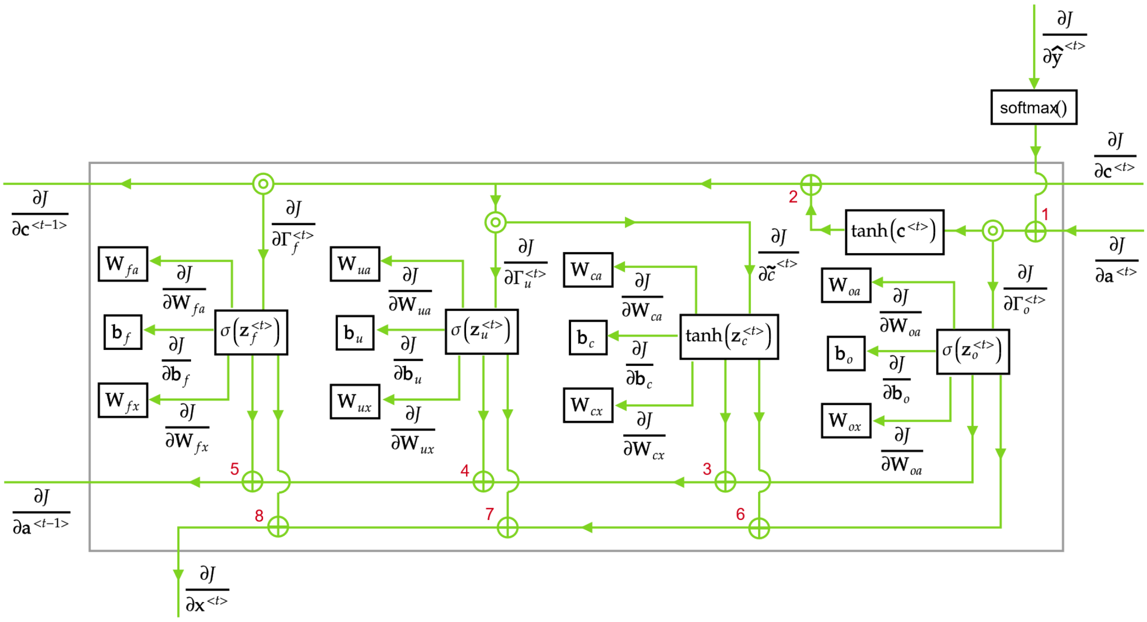

Figure 2: backward-propagation/gradient-flow diagram for a LSTM-block at time-step t. The concentric circles represent a hadamard-product (i.e. $\mathbf{x} \circ \mathbf{y}$) of the input vectors.

2.1.1 Computing $\frac{\partial J}{\partial \hat{\mathbf{y}}^{\langle t \rangle}}$

Since, $\hat{\mathbf{y}}^{\langle t \rangle}$ is computed using the softmax() activation-function, the loss J is computed using the cross-entropy loss. Also, let $\mathbf{y}^{\langle t \rangle}$ be the output-label corresponding to the t^{th} time-step. Then,

$$\begin{aligned}\frac{\partial J}{\partial \hat{\mathbf{y}}^{\langle t \rangle}} &= \begin{bmatrix} \frac{\partial J}{\partial \hat{y}_1^{\langle t \rangle}} & \frac{\partial J}{\partial \hat{y}_2^{\langle t \rangle}} & \cdots & \frac{\partial J}{\partial \hat{y}_{ny}^{\langle t \rangle}} \end{bmatrix} \\ &= \begin{bmatrix} \frac{y_1^{\langle t \rangle}}{\hat{y}_1^{\langle t \rangle}} & \frac{y_2^{\langle t \rangle}}{\hat{y}_2^{\langle t \rangle}} & \cdots & \frac{y_{ny}^{\langle t \rangle}}{\hat{y}_{ny}^{\langle t \rangle}} \end{bmatrix}\end{aligned}$$

2.1.2 Computing $\frac{\partial J}{\partial \mathbf{a}^{\langle t \rangle}}$

This derivative has two components

Comp-1 one flows-in from the $(t+1)^{th}$ time-step, and

Comp-2 one flows-in as the derivative from $\hat{\mathbf{y}}^{\langle t \rangle}$. This derivative is computed as follows,

$$\frac{\partial J}{\partial \mathbf{a}^{\langle t \rangle}} = \frac{\partial J}{\partial \hat{\mathbf{y}}^{\langle t \rangle}} \frac{\partial \hat{\mathbf{y}}^{\langle t \rangle}}{\partial \mathbf{a}^{\langle t \rangle}}$$

Then, these two components are added to compute the true derivative (see the $\oplus$ labeled as *1* in the figure-2), i.e.

$$\frac{\partial J}{\partial \mathbf{a}^{\langle t \rangle}} = \left.\frac{\partial J}{\partial \mathbf{a}^{\langle t \rangle}}\right|_{\textbf{Comp-1}} + \left.\frac{\partial J}{\partial \mathbf{a}^{\langle t \rangle}}\right|_{\textbf{Comp-2}}$$

2.1.3 Computing $\frac{\partial J}{\partial \mathbf{c}^{\langle t \rangle}}$, and $\frac{\partial J}{\partial \Gamma_o^{\langle t \rangle}}$

From *eq.10* (see section-2), we have

$$\begin{aligned}
\frac{\partial J}{\partial \Gamma_o^{\langle t \rangle}} &= \frac{\partial J}{\partial \mathbf{a}^{\langle t \rangle}} \frac{\partial \mathbf{a}^{\langle t \rangle}}{\partial \Gamma_o^{\langle t \rangle}} \\
&= \frac{\partial J}{\partial \mathbf{a}^{\langle t \rangle}} \operatorname{diag}\left(\tanh(\mathbf{c}^{\langle t \rangle})\right) \\
&= \frac{\partial J}{\partial \mathbf{a}^{\langle t \rangle}} \circ \left(\tanh(\mathbf{c}^{\langle t \rangle})\right)^\intercal
\end{aligned}$$

Note: For more information about computing the derivative of a Hadamard-product, see section-4.

The derivative $\frac{\partial J}{\partial \mathbf{c}^{\langle t \rangle}}$ is made of two components, described as follows

Comp-1 one flows-in from the $(t+1)^{th}$ time-step, and

Comp-2 one flows-in as the derivative from the $\mathbf{a}^{\langle t \rangle}$. This derivative is computed as follows,

$$\begin{aligned}
\frac{\partial J}{\partial \mathbf{c}^{\langle t \rangle}} &= \frac{\partial J}{\partial \mathbf{a}^{\langle t \rangle}} \frac{\partial \mathbf{a}^{\langle t \rangle}}{\partial \mathbf{c}^{\langle t \rangle}} \\
&= \frac{\partial J}{\partial \mathbf{a}^{\langle t \rangle}} \frac{\partial \mathbf{a}^{\langle t \rangle}}{\partial \tanh(\mathbf{c}^{\langle t \rangle})} \frac{\partial \tanh(\mathbf{c}^{\langle t \rangle})}{\partial \mathbf{c}^{\langle t \rangle}} \\
&= \frac{\partial J}{\partial \mathbf{a}^{\langle t \rangle}} (\mathbf{I}_{n_a \times n_a} \circ \Gamma_o^{\langle t \rangle})(\mathbf{I}_{n_a \times n_a} - \mathbf{I}_{n_a \times n_a} \circ \tanh^2(\mathbf{c}^{\langle t \rangle})) \\
&= \frac{\partial J}{\partial \mathbf{a}^{\langle t \rangle}} \circ \left[\Gamma_o^{\langle t \rangle} \circ (\mathbf{1}_{(n_a,1)} - \tanh^2(\mathbf{c}^{\langle t \rangle})\right]^\intercal
\end{aligned}$$

These two components are then added to compute the true derivative (see $\oplus$ labeled as *2* in figure-2), i.e.

$$\frac{\partial J}{\partial \mathbf{c}^{\langle t \rangle}} = \left.\frac{\partial J}{\partial \mathbf{c}^{\langle t \rangle}}\right|_{\textbf{Comp-1}} + \left.\frac{\partial J}{\partial \mathbf{c}^{\langle t \rangle}}\right|_{\textbf{Comp-2}}$$

2.1.4 Computing $\frac{\partial J}{\partial \mathbf{W}_{oa}}$, $\frac{\partial J}{\partial \mathbf{W}_{ox}}$, and $\frac{\partial J}{\partial \mathbf{b}_o}$

From *eq.6* (see section-2), we have

$$\begin{aligned}
\frac{\partial J}{\partial \mathbf{z}_o^{\langle t \rangle}} &= \frac{\partial J}{\partial \Gamma_o^{\langle t \rangle}} \frac{\partial \Gamma_o^{\langle t \rangle}}{\partial \mathbf{z}_o^{\langle t \rangle}} \\
&= \frac{\partial J}{\partial \Gamma_o^{\langle t \rangle}} \circ \left(\Gamma_o^{\langle t \rangle} \circ (\mathbf{1}_{(n_a,1)} - \Gamma_o^{\langle t \rangle})\right)^\intercal
\end{aligned}$$

From *eq.5* (see section-2), using the trace-method, we have

$$\operatorname{tr}(J) = \operatorname{tr}\left(\frac{\partial J}{\partial \mathbf{z}_o^{\langle t \rangle}} \, \mathrm{d}\mathbf{z}_o^{\langle t \rangle}\right)$$

where, $\mathrm{d}\mathbf{z}_o^{\langle t \rangle}$ can be expanded as follows:

$$
\begin{aligned}
\mathrm{d}\mathbf{z}_o^{\langle t \rangle} &= \mathrm{d}\left(\mathbf{W}_{oa}^{\intercal}\,\mathbf{a}^{\langle t-1 \rangle} + \mathbf{W}_{ox}^{\intercal}\,\mathbf{x}^{\langle t \rangle} + \mathbf{b}_o\right) \\
&= \mathrm{d}\mathbf{W}_{oa}^{\intercal}\,\mathbf{a}^{\langle t-1 \rangle} + \mathbf{W}_{oa}^{\intercal}\,\mathrm{d}\mathbf{a}^{\langle t-1 \rangle} + \mathrm{d}\mathbf{W}_{ox}^{\intercal}\,\mathbf{x}^{\langle t \rangle} + \mathbf{W}_{ox}^{\intercal}\,\mathrm{d}\mathbf{x}^{\langle t \rangle} + \mathrm{d}\mathbf{b}_o \qquad \text{eq.12-1}
\end{aligned}
$$

when differentiating w.r.t. $\mathbf{W}_{oa}$, we have $\mathrm{d}\mathbf{W}_{ox} = 0$, $\mathrm{d}\mathbf{a}^{\langle t-1 \rangle} = 0$, $\mathrm{d}\mathbf{b}_o = 0$, and $\mathrm{d}\mathbf{x}^{\langle t \rangle} = 0$. So,

$$
\begin{aligned}
\mathrm{d}J &= \mathrm{tr}\left(\frac{\partial J}{\partial \mathbf{z}_o^{\langle t \rangle}}\,\mathrm{d}\mathbf{W}_{oa}^{\intercal}\,\mathbf{a}^{\langle t-1 \rangle}\right) \\
&= \mathrm{tr}\left(\left(\mathbf{a}^{\langle t-1 \rangle}\right)^{\intercal}\,\mathrm{d}\mathbf{W}_{oa}\left(\frac{\partial J}{\partial \mathbf{z}_o^{\langle t \rangle}}\right)^{\intercal}\right) \\
&= \mathrm{tr}\left(\left(\frac{\partial J}{\partial \mathbf{z}_o^{\langle t \rangle}}\right)^{\intercal}\left(\mathbf{a}^{\langle t-1 \rangle}\right)^{\intercal}\,\mathrm{d}\mathbf{W}_{oa}\right) \\
\Longrightarrow \frac{\partial J}{\partial \mathbf{W}_{oa}} &= \left(\frac{\partial J}{\partial \mathbf{z}_o^{\langle t \rangle}}\right)^{\intercal}\left(\mathbf{a}^{\langle t-1 \rangle}\right)^{\intercal}
\end{aligned}
$$

when differentiating w.r.t. $\mathbf{W}_{ox}$, we have $\mathrm{d}\mathbf{W}_{oa} = 0$, $\mathrm{d}\mathbf{a}^{\langle t-1 \rangle} = 0$, $\mathrm{d}\mathbf{b}_o = 0$, and $\mathrm{d}\mathbf{x}^{\langle t \rangle} = 0$. So,

$$
\begin{aligned}
\mathrm{d}J &= \mathrm{tr}\left(\frac{\partial J}{\partial \mathbf{z}_o^{\langle t \rangle}}\,\mathrm{d}\mathbf{W}_{ox}^{\intercal}\,\mathbf{x}^{\langle t \rangle}\right) \\
&= \mathrm{tr}\left(\left(\mathbf{x}^{\langle t \rangle}\right)^{\intercal}\,\mathrm{d}\mathbf{W}_{ox}\left(\frac{\partial J}{\partial \mathbf{z}_o^{\langle t \rangle}}\right)^{\intercal}\right) \\
&= \mathrm{tr}\left(\left(\frac{\partial J}{\partial \mathbf{z}_o^{\langle t \rangle}}\right)^{\intercal}\left(\mathbf{x}^{\langle t \rangle}\right)^{\intercal}\,\mathrm{d}\mathbf{W}_{ox}\right) \\
\Longrightarrow \frac{\partial J}{\partial \mathbf{W}_{ox}} &= \left(\frac{\partial J}{\partial \mathbf{z}_o^{\langle t \rangle}}\right)^{\intercal}\left(\mathbf{x}^{\langle t \rangle}\right)^{\intercal}
\end{aligned}
$$

when differentiating w.r.t. $\mathbf{b}_o$, we have $\mathrm{d}\mathbf{W}_{oa} = 0$, $\mathrm{d}\mathbf{a}^{\langle t-1 \rangle} = 0$, $\mathrm{d}\mathbf{W}_{ox} = 0$, and $\mathrm{d}\mathbf{x}^{\langle t \rangle} = 0$. So,

$$
\begin{aligned}
\mathrm{d}J &= \mathrm{tr}\left(\frac{\partial J}{\partial \mathbf{z}_o^{\langle t \rangle}}\,\mathrm{d}\mathbf{b}_o\right) \\
\Longrightarrow \frac{\partial J}{\partial \mathbf{b}_o} &= \frac{\partial J}{\partial \mathbf{z}_o^{\langle t \rangle}}
\end{aligned}
$$

2.1.5 Computing $\frac{\partial J}{\partial \Gamma_u^{\langle t \rangle}}$, $\frac{\partial J}{\partial \tilde{\mathbf{c}}^{\langle t \rangle}}$, $\frac{\partial J}{\partial \Gamma_f^{\langle t \rangle}}$, and $\frac{\partial J}{\partial \mathbf{c}^{\langle t-1 \rangle}}$

From *eq.9* (see section-2), we have

$$
\begin{aligned}
\frac{\partial J}{\partial \Gamma_u^{\langle t \rangle}} &= \frac{\partial J}{\partial \mathbf{c}^{\langle t \rangle}}\frac{\partial \mathbf{c}^{\langle t \rangle}}{\partial \Gamma_u^{\langle t \rangle}} \\
&= \frac{\partial J}{\partial \mathbf{c}^{\langle t \rangle}}\,\mathrm{diag}\left(\tilde{\mathbf{c}}^{\langle t \rangle}\right) \\
&= \frac{\partial J}{\partial \mathbf{c}^{\langle t \rangle}} \circ \left(\tilde{\mathbf{c}}^{\langle t \rangle}\right)^{\intercal}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial J}{\partial \tilde{\mathbf{c}}^{\langle t\rangle}} &= \frac{\partial J}{\partial \mathbf{c}^{\langle t\rangle}} \frac{\partial \mathbf{c}^{\langle t\rangle}}{\partial \tilde{\mathbf{c}}^{\langle t\rangle}} \\
&= \frac{\partial J}{\partial \mathbf{c}^{\langle t\rangle}} \operatorname{diag}\left(\Gamma_u^{\langle t\rangle}\right) \\
&= \frac{\partial J}{\partial \mathbf{c}^{\langle t\rangle}} \circ \left(\Gamma_u^{\langle t\rangle}\right)^\intercal
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial J}{\partial \Gamma_f^{\langle t\rangle}} &= \frac{\partial J}{\partial \mathbf{c}^{\langle t\rangle}} \frac{\partial \mathbf{c}^{\langle t\rangle}}{\partial \Gamma_f^{\langle t\rangle}} \\
&= \frac{\partial J}{\partial \mathbf{c}^{\langle t\rangle}} \operatorname{diag}\left(\mathbf{c}^{\langle t-1\rangle}\right) \\
&= \frac{\partial J}{\partial \mathbf{c}^{\langle t\rangle}} \circ \left(\mathbf{c}^{\langle t-1\rangle}\right)^\intercal
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial J}{\partial \mathbf{c}^{\langle t-1\rangle}} &= \frac{\partial J}{\partial \mathbf{c}^{\langle t\rangle}} \frac{\partial \mathbf{c}^{\langle t\rangle}}{\partial \mathbf{c}^{\langle t-1\rangle}} \\
&= \frac{\partial J}{\partial \mathbf{c}^{\langle t\rangle}} \operatorname{diag}\left(\Gamma_f^{\langle t\rangle}\right) \\
&= \frac{\partial J}{\partial \mathbf{c}^{\langle t\rangle}} \circ \left(\Gamma_f^{\langle t\rangle}\right)^\intercal
\end{aligned}
$$

2.1.6 Computing $\frac{\partial J}{\partial \mathbf{W}_{ua}}$, $\frac{\partial J}{\partial \mathbf{W}_{ux}}$, and $\frac{\partial J}{\partial \mathbf{b}_u}$

From *eq.4* (see section-2), we have

$$
\begin{aligned}
\frac{\partial J}{\partial \mathbf{z}_u^{\langle t\rangle}} &= \frac{\partial J}{\partial \Gamma_u^{\langle t\rangle}} \frac{\partial \Gamma_u^{\langle t\rangle}}{\partial \mathbf{z}_u^{\langle t\rangle}} \\
&= \frac{\partial J}{\partial \Gamma_u^{\langle t\rangle}} \circ \left(\Gamma_u^{\langle t\rangle} \circ (\mathbf{1}_{(n_a,1)} - \Gamma_u^{\langle t\rangle})\right)^\intercal
\end{aligned}
$$

From *eq.3* (see section-2), using the trace-method, we have

$$
\operatorname{tr}(J) = \operatorname{tr}\left(\frac{\partial J}{\partial \mathbf{z}_u^{\langle t\rangle}} \, \mathrm{d}\mathbf{z}_u^{\langle t\rangle}\right)
$$

where, $\mathrm{d}\mathbf{z}_u^{\langle t\rangle}$ can be expanded as follows:

$$
\begin{aligned}
\mathrm{d}\mathbf{z}_u^{\langle t\rangle} &= \mathrm{d}\left(\mathbf{W}_{ua}^\intercal \, \mathbf{a}^{\langle t-1\rangle} + \mathbf{W}_{ux}^\intercal \, \mathbf{x}^{\langle t\rangle} + \mathbf{b}_u\right) \\
&= \mathrm{d}\mathbf{W}_{ua}^\intercal \, \mathbf{a}^{\langle t-1\rangle} + \mathbf{W}_{ua}^\intercal \, \mathrm{d}\mathbf{a}^{\langle t-1\rangle} + \mathrm{d}\mathbf{W}_{ux}^\intercal \, \mathbf{x}^{\langle t\rangle} + \mathbf{W}_{ux}^\intercal \, \mathrm{d}\mathbf{x}^{\langle t\rangle} + \mathrm{d}\mathbf{b}_u \qquad \text{eq.12-2}
\end{aligned}
$$

when differentiating w.r.t. $\mathbf{W}_{ua}$, we have $\mathrm{d}\mathbf{W}_{ux} = 0$, $\mathrm{d}\mathbf{a}^{\langle t-1\rangle} = 0$, $\mathrm{d}\mathbf{b}_u = 0$, and $\mathrm{d}\mathbf{x}^{\langle t\rangle} = 0$. So,

$$
\begin{aligned}
\mathrm{d}J &= \operatorname{tr}\left(\frac{\partial J}{\partial \mathbf{z}_u^{\langle t\rangle}}\, \mathrm{d}\mathbf{W}_{ua}^{\intercal}\, \mathbf{a}^{\langle t-1\rangle}\right) \\
&= \operatorname{tr}\left(\left(\mathbf{a}^{\langle t-1\rangle}\right)^{\intercal} \mathrm{d}\mathbf{W}_{ua} \left(\frac{\partial J}{\partial \mathbf{z}_u^{\langle t\rangle}}\right)^{\intercal}\right) \\
&= \operatorname{tr}\left(\left(\frac{\partial J}{\partial \mathbf{z}_u^{\langle t\rangle}}\right)^{\intercal} \left(\mathbf{a}^{\langle t-1\rangle}\right)^{\intercal} \mathrm{d}\mathbf{W}_{ua}\right) \\
\implies \frac{\partial J}{\partial \mathbf{W}_{ua}} &= \left(\frac{\partial J}{\partial \mathbf{z}_u^{\langle t\rangle}}\right)^{\intercal} \left(\mathbf{a}^{\langle t-1\rangle}\right)^{\intercal}
\end{aligned}
$$

when differentiating w.r.t. $\mathbf{W}_{ux}$, we have $\mathrm{d}\mathbf{W}_{ua} = 0$, $\mathrm{d}\mathbf{a}^{\langle t-1\rangle} = 0$, $\mathrm{d}\mathbf{b}_u = 0$, and $\mathrm{d}\mathbf{x}^{\langle t\rangle} = 0$. So,

$$
\begin{aligned}
\mathrm{d}J &= \operatorname{tr}\left(\frac{\partial J}{\partial \mathbf{z}_u^{\langle t\rangle}}\, \mathrm{d}\mathbf{W}_{ux}^{\intercal}\, \mathbf{x}^{\langle t\rangle}\right) \\
&= \operatorname{tr}\left(\left(\mathbf{x}^{\langle t\rangle}\right)^{\intercal} \mathrm{d}\mathbf{W}_{ux} \left(\frac{\partial J}{\partial \mathbf{z}_u^{\langle t\rangle}}\right)^{\intercal}\right) \\
&= \operatorname{tr}\left(\left(\frac{\partial J}{\partial \mathbf{z}_u^{\langle t\rangle}}\right)^{\intercal} \left(\mathbf{x}^{\langle t\rangle}\right)^{\intercal} \mathrm{d}\mathbf{W}_{ux}\right) \\
\implies \frac{\partial J}{\partial \mathbf{W}_{ux}} &= \left(\frac{\partial J}{\partial \mathbf{z}_u^{\langle t\rangle}}\right)^{\intercal} \left(\mathbf{x}^{\langle t\rangle}\right)^{\intercal}
\end{aligned}
$$

when differentiating w.r.t. $\mathbf{b}_u$, we have $\mathrm{d}\mathbf{W}_{ua} = 0$, $\mathrm{d}\mathbf{a}^{\langle t-1\rangle} = 0$, $\mathrm{d}\mathbf{W}_{ux} = 0$, and $\mathrm{d}\mathbf{x}^{\langle t\rangle} = 0$. So,

$$
\begin{aligned}
\mathrm{d}J &= \operatorname{tr}\left(\frac{\partial J}{\partial \mathbf{z}_u^{\langle t\rangle}}\, \mathrm{d}\mathbf{b}_u\right) \\
\implies \frac{\partial J}{\partial \mathbf{b}_u} &= \frac{\partial J}{\partial \mathbf{z}_u^{\langle t\rangle}}
\end{aligned}
$$

2.1.7 Computing $\frac{\partial J}{\partial \mathbf{W}_{fa}}$, $\frac{\partial J}{\partial \mathbf{W}_{fx}}$, and $\frac{\partial J}{\partial \mathbf{b}_f}$

From *eq.2* (see section-2), we have

$$
\begin{aligned}
\frac{\partial J}{\partial \mathbf{z}_f^{\langle t\rangle}} &= \frac{\partial J}{\partial \Gamma_f^{\langle t\rangle}} \frac{\partial \Gamma_f^{\langle t\rangle}}{\partial \mathbf{z}_f^{\langle t\rangle}} \\
&= \frac{\partial J}{\partial \Gamma_f^{\langle t\rangle}} \circ \left(\Gamma_f^{\langle t\rangle} \circ (\mathbf{1}_{(n_a,1)} - \Gamma_f^{\langle t\rangle})\right)^{\intercal}
\end{aligned}
$$

From *eq.1* (see section-2), using the trace-method, we have

$$
\operatorname{tr}(J) = \operatorname{tr}\left(\frac{\partial J}{\partial \mathbf{z}_f^{\langle t\rangle}}\, \mathrm{d}\mathbf{z}_f^{\langle t\rangle}\right)
$$

where, $\mathrm{d}\mathbf{z}_f^{\langle t\rangle}$ can be expanded as follows:

$$
\begin{aligned}
\mathrm{d}\mathbf{z}_f^{\langle t\rangle} &= \mathrm{d}\left(\mathbf{W}_{fa}^{\intercal}\, \mathbf{a}^{\langle t-1\rangle} + \mathbf{W}_{fx}^{\intercal}\, \mathbf{x}^{\langle t\rangle} + \mathbf{b}_f\right) \\
&= \mathrm{d}\mathbf{W}_{fa}^{\intercal}\, \mathbf{a}^{\langle t-1\rangle} + \mathbf{W}_{fa}^{\intercal}\, \mathrm{d}\mathbf{a}^{\langle t-1\rangle} + \mathrm{d}\mathbf{W}_{fx}^{\intercal}\, \mathbf{x}^{\langle t\rangle} + \mathbf{W}_{fx}^{\intercal}\, \mathrm{d}\mathbf{x}^{\langle t\rangle} + \mathrm{d}\mathbf{b}_f \qquad \text{eq.12-3}
\end{aligned}
$$

when differentiating w.r.t. $\mathbf{W}_{fa}$, we have $\mathrm{d}\mathbf{W}_{fx} = 0$, $\mathrm{d}\mathbf{a}^{\langle t-1 \rangle} = 0$, $\mathrm{d}\mathbf{b}_f = 0$, and $\mathrm{d}\mathbf{x}^{\langle t \rangle} = 0$. So,

$$\begin{aligned}
\mathrm{d}J &= \mathrm{tr}\left(\frac{\partial J}{\partial \mathbf{z}_f^{\langle t \rangle}}\, \mathrm{d}\mathbf{W}_{fa}^{\intercal}\, \mathbf{a}^{\langle t-1 \rangle}\right) \\
&= \mathrm{tr}\left(\left(\mathbf{a}^{\langle t-1 \rangle}\right)^{\intercal} \mathrm{d}\mathbf{W}_{fa} \left(\frac{\partial J}{\partial \mathbf{z}_f^{\langle t \rangle}}\right)^{\intercal}\right) \\
&= \mathrm{tr}\left(\left(\frac{\partial J}{\partial \mathbf{z}_f^{\langle t \rangle}}\right)^{\intercal} \left(\mathbf{a}^{\langle t-1 \rangle}\right)^{\intercal} \mathrm{d}\mathbf{W}_{fa}\right) \\
\implies \frac{\partial J}{\partial \mathbf{W}_{fa}} &= \left(\frac{\partial J}{\partial \mathbf{z}_f^{\langle t \rangle}}\right)^{\intercal} \left(\mathbf{a}^{\langle t-1 \rangle}\right)^{\intercal}
\end{aligned}$$

when differentiating w.r.t. $\mathbf{W}_{fx}$, we have $\mathrm{d}\mathbf{W}_{fa} = 0$, $\mathrm{d}\mathbf{a}^{\langle t-1 \rangle} = 0$, $\mathrm{d}\mathbf{b}_f = 0$, and $\mathrm{d}\mathbf{x}^{\langle t \rangle} = 0$. So,

$$\begin{aligned}
\mathrm{d}J &= \mathrm{tr}\left(\frac{\partial J}{\partial \mathbf{z}_f^{\langle t \rangle}}\, \mathrm{d}\mathbf{W}_{fx}^{\intercal}\, \mathbf{x}^{\langle t \rangle}\right) \\
&= \mathrm{tr}\left(\left(\mathbf{x}^{\langle t \rangle}\right)^{\intercal} \mathrm{d}\mathbf{W}_{fx} \left(\frac{\partial J}{\partial \mathbf{z}_f^{\langle t \rangle}}\right)^{\intercal}\right) \\
&= \mathrm{tr}\left(\left(\frac{\partial J}{\partial \mathbf{z}_f^{\langle t \rangle}}\right)^{\intercal} \left(\mathbf{x}^{\langle t \rangle}\right)^{\intercal} \mathrm{d}\mathbf{W}_{fx}\right) \\
\implies \frac{\partial J}{\partial \mathbf{W}_{fx}} &= \left(\frac{\partial J}{\partial \mathbf{z}_f^{\langle t \rangle}}\right)^{\intercal} \left(\mathbf{x}^{\langle t \rangle}\right)^{\intercal}
\end{aligned}$$

when differentiating w.r.t. $\mathbf{b}_f$, we have $\mathrm{d}\mathbf{W}_{fa} = 0$, $\mathrm{d}\mathbf{a}^{\langle t-1 \rangle} = 0$, $\mathrm{d}\mathbf{W}_{fx} = 0$, and $\mathrm{d}\mathbf{x}^{\langle t \rangle} = 0$. So,

$$\begin{aligned}
\mathrm{d}J &= \mathrm{tr}\left(\frac{\partial J}{\partial \mathbf{z}_f^{\langle t \rangle}}\, \mathrm{d}\mathbf{b}_f\right) \\
\implies \frac{\partial J}{\partial \mathbf{b}_f} &= \frac{\partial J}{\partial \mathbf{z}_f^{\langle t \rangle}}
\end{aligned}$$

2.1.8 Computing $\frac{\partial J}{\partial \mathbf{W}_{ca}}$, $\frac{\partial J}{\partial \mathbf{W}_{cx}}$, and $\frac{\partial J}{\partial \mathbf{b}_c}$

From *eq.8* (see section-2), we have

$$\begin{aligned}
\frac{\partial J}{\partial \mathbf{z}_c^{\langle t \rangle}} &= \frac{\partial J}{\partial \tilde{\mathbf{c}}^{\langle t \rangle}} \frac{\partial \tilde{\mathbf{c}}^{\langle t \rangle}}{\partial \mathbf{z}_c^{\langle t \rangle}} \\
&= \frac{\partial J}{\partial \tilde{\mathbf{c}}^{\langle t \rangle}} \circ \left[\mathbf{1}_{(n_a,1)} - \tanh^2(\mathbf{z}_c^{\langle t \rangle})\right]^{\intercal}
\end{aligned}$$

From *eq.7* (see section-2), using the trace-method, we have

$$\mathrm{tr}(J) = \mathrm{tr}\left(\frac{\partial J}{\partial \mathbf{z}_c^{\langle t\rangle}}\, \mathrm{d}\mathbf{z}_c^{\langle t\rangle}\right)$$

where, $\mathrm{d}\mathbf{z}_c^{\langle t\rangle}$ can be expanded as follows:

$$\begin{aligned}
\mathrm{d}\mathbf{z}_c^{\langle t\rangle} &= \mathrm{d}\left(\mathbf{W}_{ca}^{\intercal}\,\mathbf{a}^{\langle t-1\rangle} + \mathbf{W}_{cx}^{\intercal}\,\mathbf{x}^{\langle t\rangle} + \mathbf{b}_c\right) \\
&= \mathrm{d}\mathbf{W}_{ca}^{\intercal}\,\mathbf{a}^{\langle t-1\rangle} + \mathbf{W}_{ca}^{\intercal}\,\mathrm{d}\mathbf{a}^{\langle t-1\rangle} + \mathrm{d}\mathbf{W}_{cx}^{\intercal}\,\mathbf{x}^{\langle t\rangle} + \mathbf{W}_{cx}^{\intercal}\,\mathrm{d}\mathbf{x}^{\langle t\rangle} + \mathrm{d}\mathbf{b}_c \qquad \text{eq.12-4}
\end{aligned}$$

when differentiating w.r.t. $\mathbf{W}_{ca}$, we have $\mathrm{d}\mathbf{W}_{cx} = 0$, $\mathrm{d}\mathbf{a}^{\langle t-1\rangle} = 0$, $\mathrm{d}\mathbf{b}_c = 0$, and $\mathrm{d}\mathbf{x}^{\langle t\rangle} = 0$. So,

$$\begin{aligned}
\mathrm{d}J &= \mathrm{tr}\left(\frac{\partial J}{\partial \mathbf{z}_c^{\langle t\rangle}}\,\mathrm{d}\mathbf{W}_{ca}^{\intercal}\,\mathbf{a}^{\langle t-1\rangle}\right) \\
&= \mathrm{tr}\left(\left(\mathbf{a}^{\langle t-1\rangle}\right)^{\intercal}\,\mathrm{d}\mathbf{W}_{ca}\left(\frac{\partial J}{\partial \mathbf{z}_c^{\langle t\rangle}}\right)^{\intercal}\right) \\
&= \mathrm{tr}\left(\left(\frac{\partial J}{\partial \mathbf{z}_c^{\langle t\rangle}}\right)^{\intercal}\left(\mathbf{a}^{\langle t-1\rangle}\right)^{\intercal}\,\mathrm{d}\mathbf{W}_{ca}\right) \\
\Longrightarrow \frac{\partial J}{\partial \mathbf{W}_{ca}} &= \left(\frac{\partial J}{\partial \mathbf{z}_c^{\langle t\rangle}}\right)^{\intercal}\left(\mathbf{a}^{\langle t-1\rangle}\right)^{\intercal}
\end{aligned}$$

when differentiating w.r.t. $\mathbf{W}_{cx}$, we have $\mathrm{d}\mathbf{W}_{ca} = 0$, $\mathrm{d}\mathbf{a}^{\langle t-1\rangle} = 0$, $\mathrm{d}\mathbf{b}_c = 0$, and $\mathrm{d}\mathbf{x}^{\langle t\rangle} = 0$. So,

$$\begin{aligned}
\mathrm{d}J &= \mathrm{tr}\left(\frac{\partial J}{\partial \mathbf{z}_c^{\langle t\rangle}}\,\mathrm{d}\mathbf{W}_{cx}^{\intercal}\,\mathbf{x}^{\langle t\rangle}\right) \\
&= \mathrm{tr}\left(\left(\mathbf{x}^{\langle t\rangle}\right)^{\intercal}\,\mathrm{d}\mathbf{W}_{cx}\left(\frac{\partial J}{\partial \mathbf{z}_c^{\langle t\rangle}}\right)^{\intercal}\right) \\
&= \mathrm{tr}\left(\left(\frac{\partial J}{\partial \mathbf{z}_c^{\langle t\rangle}}\right)^{\intercal}\left(\mathbf{x}^{\langle t\rangle}\right)^{\intercal}\,\mathrm{d}\mathbf{W}_{cx}\right) \\
\Longrightarrow \frac{\partial J}{\partial \mathbf{W}_{cx}} &= \left(\frac{\partial J}{\partial \mathbf{z}_c^{\langle t\rangle}}\right)^{\intercal}\left(\mathbf{x}^{\langle t\rangle}\right)^{\intercal}
\end{aligned}$$

when differentiating w.r.t. $\mathbf{b}_c$, we have $\mathrm{d}\mathbf{W}_{ca} = 0$, $\mathrm{d}\mathbf{a}^{\langle t-1\rangle} = 0$, $\mathrm{d}\mathbf{W}_{cx} = 0$, and $\mathrm{d}\mathbf{x}^{\langle t\rangle} = 0$. So,

$$\begin{aligned}
\mathrm{d}J &= \mathrm{tr}\left(\frac{\partial J}{\partial \mathbf{z}_c^{\langle t\rangle}}\,\mathrm{d}\mathbf{b}_c\right) \\
\Longrightarrow \frac{\partial J}{\partial \mathbf{b}_c} &= \frac{\partial J}{\partial \mathbf{z}_c^{\langle t\rangle}}
\end{aligned}$$

2.1.9 Computing $\frac{\partial J}{\partial \mathbf{x}^{\langle t\rangle}}$, and $\frac{\partial J}{\partial \mathbf{a}^{\langle t-1\rangle}}$

Each of these derivatives has four components, described as follows

Comp-1 flows-in as derivative from $\Gamma_o^{\langle t\rangle}$

Comp-2 flows in as derivative from $\tilde{\mathbf{c}}^{\langle t\rangle}$

Comp-3 flows-in as derivative from $\Gamma_u^{\langle t \rangle}$

Comp-4 flows-in as derivative from $\Gamma_f^{\langle t \rangle}$

These four components are then added to compute the true derivative, i.e.

$$\frac{\partial J}{\partial \mathbf{a}^{\langle t-1 \rangle}} = \left.\frac{\partial J}{\partial \mathbf{a}^{\langle t-1 \rangle}}\right|_{\textbf{Comp-1}} + \left.\frac{\partial J}{\partial \mathbf{a}^{\langle t-1 \rangle}}\right|_{\textbf{Comp-2}} + \left.\frac{\partial J}{\partial \mathbf{a}^{\langle t-1 \rangle}}\right|_{\textbf{Comp-3}} + \left.\frac{\partial J}{\partial \mathbf{a}^{\langle t-1 \rangle}}\right|_{\textbf{Comp-4}}$$

(see $\oplus$ labeled as *3*, *4*, and *5* in figure-2)

$$\frac{\partial J}{\partial \mathbf{x}^{\langle t \rangle}} = \left.\frac{\partial J}{\partial \mathbf{x}^{\langle t \rangle}}\right|_{\textbf{Comp-1}} + \left.\frac{\partial J}{\partial \mathbf{x}^{\langle t \rangle}}\right|_{\textbf{Comp-2}} + \left.\frac{\partial J}{\partial \mathbf{x}^{\langle t \rangle}}\right|_{\textbf{Comp-3}} + \left.\frac{\partial J}{\partial \mathbf{x}^{\langle t \rangle}}\right|_{\textbf{Comp-4}}$$

(see $\oplus$ labeled as *6*, *7*, and *8* in figure-2)

3 Gradient or Jacobian?

In the above derivations, we have used the numerator layout while performing matrix-derivatives. One of the consequences of this decision is that the derivatives that we have computed are in-fact jacobians and not gradients. Fortunately, gradients are just transpose of jacobians.

4 Appendix-A:

In this section we will derive an expression for the derivative of a Hadamard product of two vectors. Let $\mathbf{x}$ and $\mathbf{y}$ be two vectors, and let $\mathbf{w} = \mathbf{x} \circ \mathbf{y}$ such that

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}; \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}; \qquad \mathbf{w} = \mathbf{x} \circ \mathbf{y} = \begin{bmatrix} x_1 y_1 \\ x_2 y_2 \\ \vdots \\ x_n y_n \end{bmatrix}$$

then, we have

$$\begin{aligned} \frac{\partial \mathbf{w}}{\partial \mathbf{x}} &= \begin{bmatrix} \frac{\partial}{\partial x_1} & \frac{\partial}{\partial x_2} & \dots & \frac{\partial}{\partial x_n} \end{bmatrix} \otimes \begin{bmatrix} x_1 y_1 \\ x_2 y_2 \\ \vdots \\ x_n y_n \end{bmatrix} \\ &= \begin{bmatrix} \frac{\partial(x_1 y_1)}{\partial x_1} & \frac{\partial(x_1 y_1)}{\partial x_2} & \dots & \frac{\partial(x_1 y_1)}{\partial x_n} \\ \frac{\partial(x_2 y_2)}{\partial x_1} & \frac{\partial(x_2 y_2)}{\partial x_2} & \dots & \frac{\partial(x_2 y_2)}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial(x_n y_n)}{\partial x_1} & \frac{\partial(x_n y_n)}{\partial x_2} & \dots & \frac{\partial(x_n y_n)}{\partial x_n} \end{bmatrix} \\ &= \begin{bmatrix} y_1 & 0 & \dots & 0 \\ 0 & y_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & y_n \end{bmatrix} \\ &= \mathbf{diag}\left(\mathbf{y}\right) \end{aligned} \tag{1}$$