Gated Recurrent Unit (GRU): back-propagation derivation

Harsha Vardhan

May 5, 2022

Abstract

This document contains derivation of the gradients for a Gated Recurrent Unit (a type of RNN-block) using back-propagation (or reverse-mode differentiation). For the implementation of the neural-network see the accompanying notebooks.
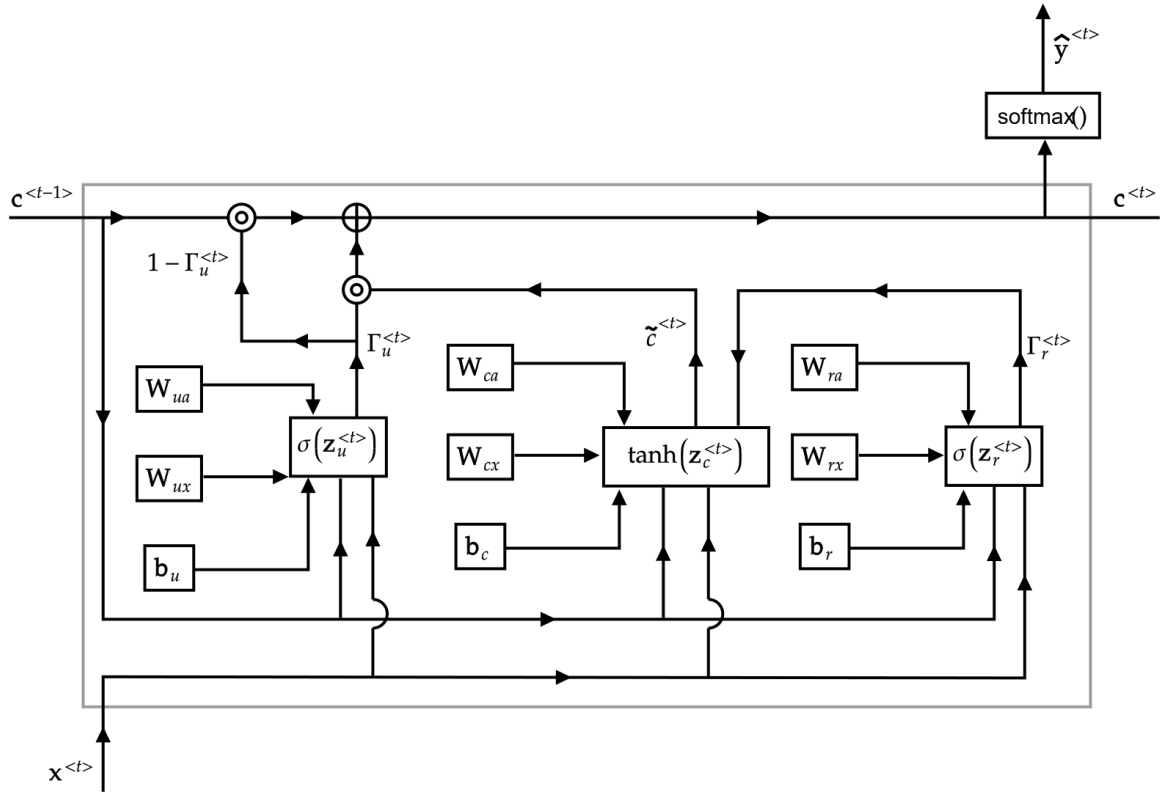
1 Block Architecture

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

$$\mathbf{z}_u^{\langle t \rangle} = \mathbf{W}_{uc}^{\mathsf{T}} \mathbf{c}^{\langle t-1 \rangle} + \mathbf{W}_{ux}^{\mathsf{T}} \mathbf{x}^{\langle t \rangle} + \mathbf{b}_u \qquad \text{eq.1}$$

$$\Gamma_u^{\langle t \rangle} = \sigma(\mathbf{z}_u^{\langle t \rangle}) \qquad \text{eq.2}$$

$$\mathbf{z}_r^{\langle t \rangle} = \mathbf{W}_{rc}^{\mathsf{T}} \mathbf{c}^{\langle t-1 \rangle} + \mathbf{W}_{rx}^{\mathsf{T}} \mathbf{x}^{\langle t \rangle} + \mathbf{b}_r \qquad \text{eq.3}$$

$$\Gamma_r^{\langle t \rangle} = \sigma(\mathbf{z}_r^{\langle t \rangle}) \qquad \text{eq.4}$$

$$\mathbf{z}_c^{\langle t \rangle} = \mathbf{W}_{cc}^{\mathsf{T}} \left(\Gamma_r^{\langle t \rangle} \circ \mathbf{c}^{\langle t-1 \rangle} \right) + \mathbf{W}_{cx}^{\mathsf{T}} \mathbf{x}^{\langle t \rangle} + \mathbf{b}_c \qquad \text{eq.5}$$

$$\tilde{\mathbf{c}}^{\langle t \rangle} = \tanh(\mathbf{z}_c^{\langle t \rangle}) \qquad \text{eq.6}$$

$$\mathbf{c}^{\langle t \rangle} = \Gamma_u^{\langle t \rangle} \circ \tilde{\mathbf{c}}^{\langle t \rangle} + (1 - \Gamma_u^{\langle t \rangle}) \circ \mathbf{c}^{\langle t-1 \rangle} \qquad \text{eq.7}$$

$$\hat{\mathbf{y}}^{\langle t \rangle} = \text{softmax}(\mathbf{c}^{\langle t \rangle}) \qquad \text{eq.8}$$

where,

$\mathbf{x}^{\langle t \rangle}$ is a $(n_x, 1)$-dimensional vector

$\mathbf{c}^{\langle t \rangle}$ & $\mathbf{c}^{\langle t-1 \rangle}$ & $\tilde{\mathbf{c}}^{\langle t \rangle}$ are $(n_a, 1)$-dimensional vectors

$\mathbf{W}_{*c}, * \in \{r, u, c\}$ are (n_a, n_a)-dimensional parameter matrices

$\mathbf{W}_{*x}, * \in \{r, u, c\}$ are (n_x, n_a)-dimensional parameter matrices

$\mathbf{b}_*, * \in \{r, u, c\}$ are $(n_a, 1)$-dimensional bias-vectors

$\Gamma_*^{\langle t \rangle}, * \in \{r, u\}$ are $(n_a, 1)$-dimensional vector, which represents a gate

$\mathbf{z}_*^{\langle t \rangle}, * \in \{r, u, c\}$ are $(n_a, 1)$-dimensional vector, is input to a gates' activation

Also, the sub-scripts denote the following: $_r$ - relevance-gate, and $_u$ - update-gate.
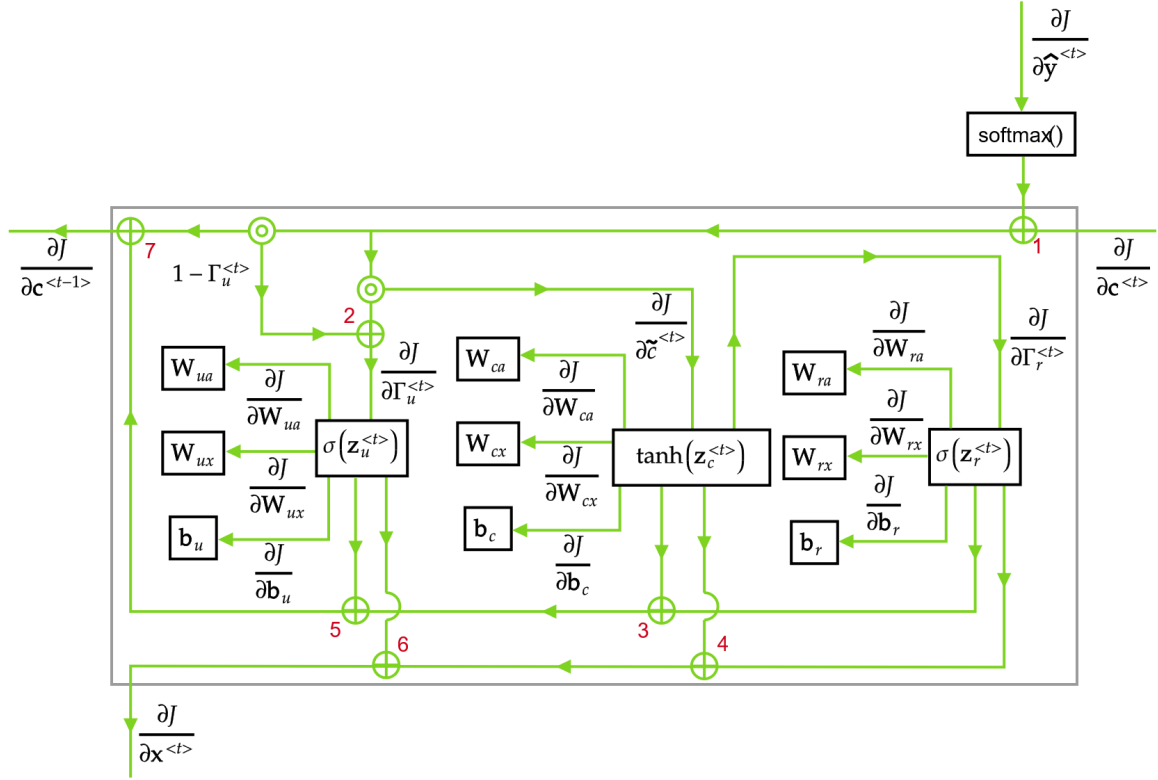
2.1 Backward Propagation

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 tth time-step. Then,

$$
\frac{\partial J}{\partial \hat{\mathbf{y}}^{\langle t \rangle}} = \begin{bmatrix} \frac{\partial J}{\partial \hat{y}_1^{\langle t \rangle}} & \frac{\partial J}{\partial \hat{y}_2^{\langle t \rangle}} & \cdots & \frac{\partial J}{\partial \hat{y}_{n_y}^{\langle t \rangle}} \end{bmatrix}
$$

$$
= \begin{bmatrix} \frac{y_1^{\langle t \rangle}}{\hat{y}_1^{\langle t \rangle}} & \frac{y_2^{\langle t \rangle}}{\hat{y}_2^{\langle t \rangle}} & \cdots & \frac{y_{n_y}^{\langle t \rangle}}{\hat{y}_{n_y}^{\langle t \rangle}} \end{bmatrix}
$$

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

This derivative has two components

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

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

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

$$
= \frac{\partial J}{\partial \hat{\mathbf{y}}^{\langle t \rangle}} \left(\mathrm{diag}(\hat{\mathbf{y}}^{\langle t \rangle}) - \hat{\mathbf{y}}^{\langle t \rangle} (\hat{\mathbf{y}}^{\langle t \rangle})^{\mathsf{T}} \right)
$$

3

Note: for more on the derivation of the derivative of a `softmax()` function, see `..\notes\softmax-function.ipynb`

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

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

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

From *eq.7* (see section-2), we have

$$\frac{\partial J}{\partial \tilde{\mathbf{c}}^{\langle t \rangle}} = \frac{\partial J}{\partial \mathbf{c}^{\langle t \rangle}} \frac{\partial \mathbf{c}^{\langle t \rangle}}{\partial \tilde{\mathbf{c}}^{\langle t \rangle}}$$

$$= \frac{\partial J}{\partial \mathbf{c}^{\langle t \rangle}} \operatorname{diag}\left(\Gamma_u^{\langle t \rangle}\right)$$

$$= \frac{\partial J}{\partial \mathbf{c}^{\langle t \rangle}} \circ \left(\Gamma_u^{\langle t \rangle}\right)^{\intercal}$$

$$\frac{\partial J}{\partial \Gamma_u^{\langle t \rangle}} = \frac{\partial J}{\partial \mathbf{c}^{\langle t \rangle}} \frac{\partial \mathbf{c}^{\langle t \rangle}}{\partial \Gamma_u^{\langle t \rangle}}$$

$$= \frac{\partial J}{\partial \mathbf{c}^{\langle t \rangle}} \operatorname{diag}\left(\tilde{\mathbf{c}}^{\langle t \rangle} - \mathbf{c}^{\langle t-1 \rangle}\right)$$

$$= \frac{\partial J}{\partial \mathbf{c}^{\langle t \rangle}} \circ \left(\tilde{\mathbf{c}}^{\langle t \rangle} - \mathbf{c}^{\langle t-1 \rangle}\right)^{\intercal}$$

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

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

$$\frac{\partial J}{\partial \mathbf{z}_c^{\langle t \rangle}} = \frac{\partial J}{\partial \tilde{\mathbf{c}}^{\langle t \rangle}} \frac{\partial \tilde{\mathbf{c}}^{\langle t \rangle}}{\partial \mathbf{z}_c^{\langle t \rangle}}$$

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

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

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

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

$$\mathrm{d}\mathbf{z}_c^{\langle t \rangle} = \mathrm{d}\left(\mathbf{W}_{cc}^{\intercal}\left(\Gamma_r^{\langle t \rangle} \circ \mathbf{c}^{\langle t-1 \rangle}\right) + \mathbf{W}_{cx}^{\intercal}\mathbf{x}^{\langle t \rangle} + \mathbf{b}_c\right)$$

$$= \mathrm{d}\mathbf{W}_{cc}^{\intercal}\left(\Gamma_r^{\langle t \rangle} \circ \mathbf{c}^{\langle t-1 \rangle}\right) + \mathbf{W}_{cc}^{\intercal}\,\mathrm{d}\left(\Gamma_r^{\langle t \rangle} \circ \mathbf{c}^{\langle t-1 \rangle}\right) + \mathrm{d}\mathbf{W}_{cx}^{\intercal}\mathbf{x}^{\langle t \rangle}$$

$$+ \mathbf{W}_{cx}^{\intercal}\,\mathrm{d}\mathbf{x}^{\langle t \rangle} + \mathrm{d}\mathbf{b}_c \qquad \text{eq.9-1}$$

when differentiating w.r.t. $\mathbf{W}_{cc}$, we have $\mathrm{d}\mathbf{W}_{cx} = 0$, $\mathrm{d}\left(\Gamma_r^{\langle t \rangle} \circ \mathbf{c}^{\langle t-1 \rangle}\right) = 0$, $\mathrm{d}\mathbf{b}_c = 0$, and $\mathrm{d}\mathbf{x}^{\langle t \rangle} = 0$. So,

$$\mathrm{d}J = \mathrm{tr}\left(\frac{\partial J}{\partial \mathbf{z}_c^{\langle t \rangle}}\, \mathrm{d}\mathbf{W}_{cc}^{\intercal}\left(\Gamma_r^{\langle t \rangle} \circ \mathbf{c}^{\langle t-1 \rangle}\right)\right)$$

$$= \mathrm{tr}\left(\left(\Gamma_r^{\langle t \rangle} \circ \mathbf{c}^{\langle t-1 \rangle}\right)^{\intercal} \mathrm{d}\mathbf{W}_{cc}\left(\frac{\partial J}{\partial \mathbf{z}_c^{\langle t \rangle}}\right)^{\intercal}\right)$$

$$= \mathrm{tr}\left(\left(\frac{\partial J}{\partial \mathbf{z}_c^{\langle t \rangle}}\right)^{\intercal}\left(\Gamma_r^{\langle t \rangle} \circ \mathbf{c}^{\langle t-1 \rangle}\right)^{\intercal} \mathrm{d}\mathbf{W}_{cc}\right)$$

$$\implies \frac{\partial J}{\partial \mathbf{W}_{cc}} = \left(\frac{\partial J}{\partial \mathbf{z}_c^{\langle t \rangle}}\right)^{\intercal}\left(\Gamma_r^{\langle t \rangle} \circ \mathbf{c}^{\langle t-1 \rangle}\right)^{\intercal}$$

when differentiating w.r.t. $\mathbf{W}_{cx}$, we have $\mathrm{d}\mathbf{W}_{cc} = 0$, $\mathrm{d}\left(\Gamma_r^{\langle t \rangle} \circ \mathbf{c}^{\langle t-1 \rangle}\right) = 0$, $\mathrm{d}\mathbf{b}_c = 0$, and $\mathrm{d}\mathbf{x}^{\langle t \rangle} = 0$. So,

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

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

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

$$\implies \frac{\partial J}{\partial \mathbf{W}_{cx}} = \left(\frac{\partial J}{\partial \mathbf{z}_c^{\langle t \rangle}}\right)^{\intercal}\left(\mathbf{x}^{\langle t \rangle}\right)^{\intercal}$$

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

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

$$\implies \frac{\partial J}{\partial \mathbf{b}_c} = \frac{\partial J}{\partial \mathbf{z}_c^{\langle t \rangle}}$$

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

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

$$\frac{\partial J}{\partial \mathbf{z}_u^{\langle t \rangle}} = \frac{\partial J}{\partial \Gamma_u^{\langle t \rangle}}\frac{\partial \Gamma_u^{\langle t \rangle}}{\partial \mathbf{z}_u^{\langle t \rangle}}$$

$$= \frac{\partial J}{\partial \Gamma_u^{\langle t \rangle}} \circ \left(\Gamma_u^{\langle t \rangle} \circ \left(\mathbf{1}_{(n_c,1)} - \Gamma_u^{\langle t \rangle}\right)\right)^{\intercal}$$

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

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

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

$$\mathrm{d}\mathbf{z}_u^{\langle t \rangle} = \mathrm{d}\left(\mathbf{W}_{uc}^{\intercal}\, \mathbf{c}^{\langle t-1 \rangle} + \mathbf{W}_{ux}^{\intercal}\, \mathbf{x}^{\langle t \rangle} + \mathbf{b}_u\right)$$

$$= \mathrm{d}\mathbf{W}_{uc}^{\intercal}\, \mathbf{c}^{\langle t-1 \rangle} + \mathbf{W}_{uc}^{\intercal}\, \mathrm{d}\mathbf{c}^{\langle t-1 \rangle} + \mathrm{d}\mathbf{W}_{ux}^{\intercal}\, \mathbf{x}^{\langle t \rangle} + \mathbf{W}_{ux}^{\intercal}\, \mathrm{d}\mathbf{x}^{\langle t \rangle} + \mathrm{d}\mathbf{b}_u \qquad \text{eq.9-2}$$

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

$$\mathrm{d}J = \mathrm{tr}\left(\frac{\partial J}{\partial \mathbf{z}_u^{\langle t \rangle}} \, \mathrm{d}\mathbf{W}_{uc}^{\mathsf{T}} \, \mathbf{c}^{\langle t-1 \rangle} \right)$$

$$= \mathrm{tr}\left(\left(\mathbf{c}^{\langle t-1 \rangle} \right)^{\mathsf{T}} \mathrm{d}\mathbf{W}_{uc} \left(\frac{\partial J}{\partial \mathbf{z}_u^{\langle t \rangle}} \right)^{\mathsf{T}} \right)$$

$$= \mathrm{tr}\left(\left(\frac{\partial J}{\partial \mathbf{z}_u^{\langle t \rangle}} \right)^{\mathsf{T}} \left(\mathbf{c}^{\langle t-1 \rangle} \right)^{\mathsf{T}} \mathrm{d}\mathbf{W}_{uc} \right)$$

$$\implies \frac{\partial J}{\partial \mathbf{W}_{uc}} = \left(\frac{\partial J}{\partial \mathbf{z}_u^{\langle t \rangle}} \right)^{\mathsf{T}} \left(\mathbf{c}^{\langle t-1 \rangle} \right)^{\mathsf{T}}$$

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

$$\mathrm{d}J = \mathrm{tr}\left(\frac{\partial J}{\partial \mathbf{z}_u^{\langle t \rangle}} \, \mathrm{d}\mathbf{W}_{ux}^{\mathsf{T}} \, \mathbf{x}^{\langle t \rangle} \right)$$

$$= \mathrm{tr}\left(\left(\mathbf{x}^{\langle t \rangle} \right)^{\mathsf{T}} \mathrm{d}\mathbf{W}_{ux} \left(\frac{\partial J}{\partial \mathbf{z}_u^{\langle t \rangle}} \right)^{\mathsf{T}} \right)$$

$$= \mathrm{tr}\left(\left(\frac{\partial J}{\partial \mathbf{z}_u^{\langle t \rangle}} \right)^{\mathsf{T}} \left(\mathbf{x}^{\langle t \rangle} \right)^{\mathsf{T}} \mathrm{d}\mathbf{W}_{ux} \right)$$

$$\implies \frac{\partial J}{\partial \mathbf{W}_{ux}} = \left(\frac{\partial J}{\partial \mathbf{z}_u^{\langle t \rangle}} \right)^{\mathsf{T}} \left(\mathbf{x}^{\langle t \rangle} \right)^{\mathsf{T}}$$

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

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

$$\implies \frac{\partial J}{\partial \mathbf{b}_u} = \frac{\partial J}{\partial \mathbf{z}_u^{\langle t \rangle}}$$

2.1.6 Computing $\frac{\partial J}{\partial \Gamma_r^{\langle t \rangle}}$

From the results in section-2.1.4 and, from *eq.9-1*, we have

$$\frac{\partial J}{\partial \Gamma_r^{\langle t \rangle}} = \frac{\partial J}{\partial \tilde{\mathbf{c}}^{\langle t \rangle}} \frac{\partial \tilde{\mathbf{c}}^{\langle t \rangle}}{\partial \Gamma_r^{\langle t \rangle}}$$

$$= \frac{\partial J}{\partial \tilde{\mathbf{c}}^{\langle t \rangle}} \frac{\partial \tilde{\mathbf{c}}^{\langle t \rangle}}{\partial \left(\Gamma_r^{\langle t \rangle} \circ \mathbf{c}^{\langle t-1 \rangle} \right)} \frac{\partial \left(\Gamma_r^{\langle t \rangle} \circ \mathbf{c}^{\langle t-1 \rangle} \right)}{\partial \Gamma_r^{\langle t \rangle}}$$

$$= \frac{\partial J}{\partial \tilde{\mathbf{c}}^{\langle t \rangle}} \mathbf{W}_{cc}^{\mathsf{T}} \, \mathrm{diag}\left(\mathbf{c}^{\langle t-1 \rangle} \right)$$

$$= \left(\frac{\partial J}{\partial \tilde{\mathbf{c}}^{\langle t \rangle}} \mathbf{W}_{cc}^{\mathsf{T}} \right) \circ \left(\mathbf{c}^{\langle t-1 \rangle} \right)^{\mathsf{T}}$$

Note: for more information on the derivative of a Hadamard product, see the Appendix-A in `.\backprop-lstm.pdf`.

2.1.7 Computing $\frac{\partial J}{\partial \mathbf{W}_{rc}}$, $\frac{\partial J}{\partial \mathbf{W}_{rx}}$, and $\frac{\partial J}{\partial \mathbf{b}_r}$

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

$$
\begin{aligned}
\frac{\partial J}{\partial \mathbf{z}_r^{\langle t \rangle}} &= \frac{\partial J}{\partial \Gamma_r^{\langle t \rangle}} \frac{\partial \Gamma_r^{\langle t \rangle}}{\partial \mathbf{z}_r^{\langle t \rangle}} \\
&= \frac{\partial J}{\partial \Gamma_r^{\langle t \rangle}} \circ \left(\Gamma_r^{\langle t \rangle} \circ \left(\mathbf{1}_{(n_c,1)} - \Gamma_r^{\langle t \rangle} \right) \right)^{\mathsf{T}}
\end{aligned}
$$

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

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

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

$$
\begin{aligned}
\mathrm{d}\mathbf{z}_r^{\langle t \rangle} &= \mathrm{d}\left(\mathbf{W}_{rc}^{\mathsf{T}} \mathbf{c}^{\langle t-1 \rangle} + \mathbf{W}_{rx}^{\mathsf{T}} \mathbf{x}^{\langle t \rangle} + \mathbf{b}_r \right) \\
&= \mathrm{d}\mathbf{W}_{rc}^{\mathsf{T}} \mathbf{c}^{\langle t-1 \rangle} + \mathbf{W}_{rc}^{\mathsf{T}} \mathrm{d}\mathbf{c}^{\langle t-1 \rangle} + \mathrm{d}\mathbf{W}_{rx}^{\mathsf{T}} \mathbf{x}^{\langle t \rangle} + \mathbf{W}_{rx}^{\mathsf{T}} \mathrm{d}\mathbf{x}^{\langle t \rangle} + \mathrm{d}\mathbf{b}_r \qquad \text{eq.9-3}
\end{aligned}
$$

when differentiating w.r.t. $\mathbf{W}_{rc}$, we have $\mathrm{d}\mathbf{W}_{rx} = 0$, $\mathrm{d}\mathbf{c}^{\langle t-1 \rangle} = 0$, $\mathrm{d}\mathbf{b}_r = 0$, and $\mathrm{d}\mathbf{x}^{\langle t \rangle} = 0$. So,

$$
\begin{aligned}
\mathrm{d}J &= \mathrm{tr}\left(\frac{\partial J}{\partial \mathbf{z}_r^{\langle t \rangle}} \, \mathrm{d}\mathbf{W}_{rc}^{\mathsf{T}} \mathbf{c}^{\langle t-1 \rangle} \right) \\
&= \mathrm{tr}\left(\left(\mathbf{c}^{\langle t-1 \rangle} \right)^{\mathsf{T}} \mathrm{d}\mathbf{W}_{rc} \left(\frac{\partial J}{\partial \mathbf{z}_r^{\langle t \rangle}} \right)^{\mathsf{T}} \right) \\
&= \mathrm{tr}\left(\left(\frac{\partial J}{\partial \mathbf{z}_r^{\langle t \rangle}} \right)^{\mathsf{T}} \left(\mathbf{c}^{\langle t-1 \rangle} \right)^{\mathsf{T}} \mathrm{d}\mathbf{W}_{rc} \right) \\
\implies \frac{\partial J}{\partial \mathbf{W}_{rc}} &= \left(\frac{\partial J}{\partial \mathbf{z}_r^{\langle t \rangle}} \right)^{\mathsf{T}} \left(\mathbf{c}^{\langle t-1 \rangle} \right)^{\mathsf{T}}
\end{aligned}
$$

when differentiating w.r.t. $\mathbf{W}_{rx}$, we have $\mathrm{d}\mathbf{W}_{rc} = 0$, $\mathrm{d}\mathbf{c}^{\langle t-1 \rangle} = 0$, $\mathrm{d}\mathbf{b}_r = 0$, and $\mathrm{d}\mathbf{x}^{\langle t \rangle} = 0$. So,

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

when differentiating w.r.t. $\mathbf{b}_r$, we have $\mathrm{d}\mathbf{W}_{rc} = 0$, $\mathrm{d}\mathbf{c}^{\langle t-1 \rangle} = 0$, $\mathrm{d}\mathbf{W}_{rx} = 0$, and $\mathrm{d}\mathbf{x}^{\langle t \rangle} = 0$. So,

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

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

The derivative $\frac{\partial J}{\partial \mathbf{c}^{\langle t-1 \rangle}}$ has four components, as follows

Comp-1 flows-in as part of the derivative $\frac{\partial J}{\partial \mathbf{c}^{\langle t \rangle}}$. This derivative is computed as follows

$$
\begin{aligned}
\frac{\partial J}{\partial \mathbf{c}^{\langle t-1 \rangle}} &= \frac{\partial J}{\partial \mathbf{c}^{\langle t \rangle}} \frac{\partial \mathbf{c}^{\langle t \rangle}}{\partial \mathbf{c}^{\langle t-1 \rangle}} \\
&= \frac{\partial J}{\partial \mathbf{c}^{\langle t \rangle}} \operatorname{diag}\left(\mathbf{1}_{(n_c,1)} - \Gamma_u^{\langle t \rangle}\right) \\
&= \frac{\partial J}{\partial \mathbf{c}^{\langle t \rangle}} \circ \left(\mathbf{1}_{(n_c,1)} - \Gamma_u^{\langle t \rangle}\right)^{\mathsf{T}}
\end{aligned}
$$

Comp-2 flows-in as derivative from $\Gamma_r^{\langle t \rangle}$ (see *eq.9-3*), and is computed as follows

$$
\begin{aligned}
\frac{\partial J}{\partial \mathbf{c}^{\langle t-1 \rangle}} &= \frac{\partial J}{\partial \mathbf{z}_r^{\langle t \rangle}} \frac{\partial \mathbf{z}_r^{\langle t \rangle}}{\partial \mathbf{c}^{\langle t-1 \rangle}} \\
&= \frac{\partial J}{\partial \mathbf{z}_r^{\langle t \rangle}} \mathbf{W}_{rc}^{\mathsf{T}}
\end{aligned}
$$

Comp-3 flows-in as derivative from $\Gamma_u^{\langle t \rangle}$ (see *eq.9-2*), and is computed as follows

$$
\begin{aligned}
\frac{\partial J}{\partial \mathbf{c}^{\langle t-1 \rangle}} &= \frac{\partial J}{\partial \mathbf{z}_u^{\langle t \rangle}} \frac{\partial \mathbf{z}_u^{\langle t \rangle}}{\partial \mathbf{c}^{\langle t-1 \rangle}} \\
&= \frac{\partial J}{\partial \mathbf{z}_r^{\langle t \rangle}} \mathbf{W}_{uc}^{\mathsf{T}}
\end{aligned}
$$

Comp-4 flows-in as derivative from $\tilde{\mathbf{c}}^{\langle t \rangle}$ (see *eq.9-1*), and is computed as follows

$$
\begin{aligned}
\frac{\partial J}{\partial \mathbf{c}^{\langle t-1 \rangle}} &= \frac{\partial J}{\partial \mathbf{z}_c^{\langle t \rangle}} \frac{\partial \mathbf{z}_c^{\langle t \rangle}}{\partial \mathbf{c}^{\langle t-1 \rangle}} \\
&= \frac{\partial J}{\partial \mathbf{z}_c^{\langle t \rangle}} \frac{\partial \mathbf{z}_c^{\langle t \rangle}}{\partial \left(\Gamma_r^{\langle t \rangle} \circ \mathbf{c}^{\langle t-1 \rangle}\right)} \frac{\partial \left(\Gamma_r^{\langle t \rangle} \circ \mathbf{c}^{\langle t-1 \rangle}\right)}{\partial \mathbf{c}^{\langle t-1 \rangle}} \\
&= \frac{\partial J}{\partial \mathbf{z}_c^{\langle t \rangle}} \mathbf{W}_{cc}^{\mathsf{T}} \operatorname{diag}\left(\Gamma_r^{\langle t \rangle}\right) \\
&= \left(\frac{\partial J}{\partial \mathbf{z}_c^{\langle t \rangle}} \mathbf{W}_{cc}^{\mathsf{T}}\right) \circ \left(\Gamma_r^{\langle t \rangle}\right)^{\mathsf{T}}
\end{aligned}
$$

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

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

(see $\oplus$ labeled as *3, 5,* and *7* in figure-2)

The derivative $\frac{\partial J}{\partial \mathbf{c}^{\langle t-1 \rangle}}$ has three components, as follows

8

Comp-1 flows-in as derivative from $\Gamma_r^{\langle t \rangle}$ (see *eq.9-3*), and is computed as follows

$$\frac{\partial J}{\partial \mathbf{x}^{\langle t \rangle}} = \frac{\partial J}{\partial \mathbf{z}_r^{\langle t \rangle}} \frac{\partial \mathbf{z}_r^{\langle t \rangle}}{\partial \mathbf{x}^{\langle t \rangle}}$$

$$= \frac{\partial J}{\partial \mathbf{z}_r^{\langle t \rangle}} \mathbf{W}_{rx}^{\intercal}$$

Comp-2 flows-in as derivative from $\Gamma_u^{\langle t \rangle}$ (see *eq.9-2*), and is computed as follows

$$\frac{\partial J}{\partial \mathbf{x}^{\langle t \rangle}} = \frac{\partial J}{\partial \mathbf{z}_u^{\langle t \rangle}} \frac{\partial \mathbf{z}_u^{\langle t \rangle}}{\partial \mathbf{x}^{\langle t \rangle}}$$

$$= \frac{\partial J}{\partial \mathbf{z}_r^{\langle t \rangle}} \mathbf{W}_{ux}^{\intercal}$$

Comp-3 flows-in as derivative from $\tilde{\mathbf{c}}^{\langle t \rangle}$ (see *eq.9-1*), and is computed as follows

$$\frac{\partial J}{\partial \mathbf{x}^{\langle t \rangle}} = \frac{\partial J}{\partial \mathbf{z}_c^{\langle t \rangle}} \frac{\partial \mathbf{z}_c^{\langle t \rangle}}{\partial \mathbf{x}^{\langle t \rangle}}$$

$$= \frac{\partial J}{\partial \mathbf{z}_c^{\langle t \rangle}} \mathbf{W}_{cx}^{\intercal}$$

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

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

(see $\oplus$ labeled as *4*, and *6* 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.