1. RNN model and backpropagation

The simple recurrent unit is given by the equations:

$$h_t = \mathrm{ReLU}(W_x x_t + W_h h_{t-1} + b_h) \tag{1}$$

$$y_t = \mathrm{ReLU}(W_y h_t + b_y) \tag{2}$$

The loss function is the sum of losses over the time steps:

$$L = \sum_{t=1}^{T} L_t(y_t) \tag{3}$$

As the loss is calculated outside the recurrent unit, the unit itself recieves the $\frac{\partial L_t}{\partial y_t}$ values as backpropagated errors from the next layer. The contributions to the derivatives w.r.t. the parameters W_y and b_y in each time step t can be calculated as:

$$\frac{\partial L_t}{\partial W_y} = \frac{\partial L_t}{\partial y_t}\frac{\partial y_t}{\partial W_y} = \left\{\frac{\partial L_t}{\partial y_t}\right\}\left\{\mathrm{ReLU}'(z_t)\ h_t\right\} = \left(\frac{\partial L_t}{\partial y_t} \odot \mathrm{ReLU}'\left(z_t\right)\right) h_t^{\mathrm{T}} \tag{4}$$

$$\frac{\partial L_t}{\partial b_y} = \frac{\partial L_t}{\partial y_t}\frac{\partial y_t}{\partial b_y} = \left\{\frac{\partial L_t}{\partial y_t}\right\}\left\{\mathrm{ReLU}'(z_t)\right\} = \left(\frac{\partial L_t}{\partial y_t} \odot \mathrm{ReLU}'\left(z_t\right)\right) \tag{5}$$

The derivatives w.r.t. W_x, W_h and b_h are:

$$\begin{aligned}\frac{\partial L_t}{\partial W_x} &= \frac{\partial L_t}{\partial y_t}\frac{\partial y_t}{\partial h_t}\frac{\partial h_t}{\partial W_x} = \left\{\frac{\partial L_t}{\partial y_t}\right\}\left\{\mathrm{ReLU}'(z_t)\, W_y\right\}\left\{\mathrm{ReLU}'(z_t^0)\left(x_t + W_h \frac{\partial h_{t-1}}{\partial W_x}\right)\right\} = \\ &= \left(W_y^{\mathrm{T}}\left(\frac{\partial L_t}{\partial y_t} \odot \mathrm{ReLU}'(z_t) \odot \mathrm{ReLU}'\left(z_t^0\right)\right)\right) x_t^{\mathrm{T}} + \left(W_h^{\mathrm{T}}\left(W_y^{\mathrm{T}}\left(\frac{\partial L_t}{\partial y_t} \odot \mathrm{ReLU}'(z_t) \odot \mathrm{ReLU}'\left(z_t^0\right)\right)\right)\right)\frac{\partial h_{t-1}}{\partial W_x}\end{aligned} \tag{6}$$

$$\begin{aligned}\frac{\partial L_t}{\partial W_h} &= \frac{\partial L_t}{\partial y_t}\frac{\partial y_t}{\partial h_t}\frac{\partial h_t}{\partial W_h} = \left\{\frac{\partial L_t}{\partial y_t}\right\}\left\{\mathrm{ReLU}'(z_t)\ W_y\right\}\left\{\mathrm{ReLU}'(z_t^0)\left(h_{t-1} + W_h \frac{\partial h_{t-1}}{\partial W_h}\right)\right\} = \\ &= \left(W_y^{\mathrm{T}}\left(\frac{\partial L_t}{\partial y_t} \odot \mathrm{ReLU}'(z_t) \odot \mathrm{ReLU}'\left(z_t^0\right)\right)\right) h_{t-1}^{\mathrm{T}} + \left(W_h^{\mathrm{T}}\left(W_y^{\mathrm{T}}\left(\frac{\partial L_t}{\partial y_t} \odot \mathrm{ReLU}'(z_t) \odot \mathrm{ReLU}'\left(z_t^0\right)\right)\right)\right)\frac{\partial h_{t-1}}{\partial W_h}\end{aligned} \tag{7}$$

$$\begin{aligned}\frac{\partial L_t}{\partial b_h} &= \frac{\partial L_t}{\partial y_t}\frac{\partial y_t}{\partial h_t}\frac{\partial h_t}{\partial b_h} = \left\{\frac{\partial L_t}{\partial y_t}\right\}\left\{\mathrm{ReLU}'(z_t)\ W_y\right\}\left\{\mathrm{ReLU}'(z_t^0)\left(1 + W_h \frac{\partial h_{t-1}}{\partial b_h}\right)\right\} = \\ &= \left(W_y^{\mathrm{T}}\left(\frac{\partial L_t}{\partial y_t} \odot \mathrm{ReLU}'(z_t) \odot \mathrm{ReLU}'\left(z_t^0\right)\right)\right) + \left(W_h^{\mathrm{T}}\left(W_y^{\mathrm{T}}\left(\frac{\partial L_t}{\partial y_t} \odot \mathrm{ReLU}'(z_t) \odot \mathrm{ReLU}'\left(z_t^0\right)\right)\right)\right)\frac{\partial h_{t-1}}{\partial b_h}\end{aligned} \tag{8}$$

Where $\odot$ denotes the elementwise (Hadamard) product.
Finally, the error backpropagated to the previous layer or unit is:

$$\begin{aligned}\frac{\partial L_t}{\partial x_t} &= \frac{\partial L_t}{\partial y_t}\frac{\partial y_t}{\partial h_t}\frac{\partial h_t}{\partial x_t} = \left\{\frac{\partial L_t}{\partial y_t}\right\}\left\{\mathrm{ReLU}'(z_t)\, W_y\right\}\left\{\mathrm{ReLU}'\left(z_t^0\right)\left(W_x + W_h \frac{\partial h_{t-1}}{\partial x_t}\right)\right\} = \\ &= W_x^{\mathrm{T}}\left(W_y^{\mathrm{T}}\left(\frac{\partial L_t}{\partial y_t} \odot \mathrm{ReLU}'(z_t) \odot \mathrm{ReLU}'\left(z_t^0\right)\right)\right) + \left(W_h^{\mathrm{T}}\left(W_y^{\mathrm{T}}\left(\frac{\partial L_t}{\partial y_t} \odot \mathrm{ReLU}'(z_t) \odot \mathrm{ReLU}'\left(z_t^0\right)\right)\right)\right)\frac{\partial h_{t-1}}{\partial x_t}\end{aligned} \tag{9}$$

Where z_t and z_t^0 are the preactivations:

$$z_t^0 = W_x x_t + W_h h_{t-1} + b_h \tag{10}$$

$$z_t = W_y h_t + b_y \tag{11}$$

Where each h_t was considered as a function of the previous h_{t-1}, and every derivative of h_{t-1} should also be calculated further, recursively. However, considering that the coefficients in front of the recursive partial derivatives is the error backpropagated through time:

$$\begin{aligned}\frac{\partial L_t}{\partial h_{t-1}} &= \frac{\partial L_t}{\partial y_t}\frac{\partial y_t}{\partial h_t}\frac{\partial h_t}{\partial h_{t-1}} = \left\{\frac{\partial L_t}{\partial y_t}\right\}\left\{\mathrm{ReLU}'(z_t)\, W_y\right\}\left\{\mathrm{ReLU}'\left(z_t^0\right) W_h\right\} = \\ &= W_h^{\mathrm{T}}\left(W_y^{\mathrm{T}}\left(\frac{\partial L_t}{\partial y_t} \odot \mathrm{ReLU}'(z_t) \odot \mathrm{ReLU}'\left(z_t^0\right)\right)\right)\end{aligned} \tag{12}$$

The terms containing the derivatives of h_{t-1} can be pushed to the previous time step $t-1$ by introducing the cumulative BPTT error, which contains the contribution from every subsequent time step:

$$\frac{\partial L_{t+}}{\partial h_t} = \frac{\partial L_t}{\partial h_t} + \frac{\partial L_{(t+1)+}}{\partial h_t} = W_y^{\mathrm{T}}\left(\frac{\partial L_t}{\partial y_t} \odot \mathrm{ReLU}'(z_t)\right) + \frac{\partial L_{(t+1)+}}{\partial h_t} \tag{13}$$

and we can use the substitution

$$\frac{\partial L_t}{\partial y_t}\frac{\partial y_t}{\partial h_t} \quad\rightarrow\quad \frac{\partial L_{t+}}{\partial h_t} \tag{14}$$

to account for the recursion of nested functions in each time step. Using the notations:

$$\delta_t = \frac{\partial L_t}{\partial y_t} \odot \mathrm{ReLU}'(z_t) \tag{15}$$

$$\delta_t^h = \frac{\partial L_{t+}}{\partial h_t} \odot \mathrm{ReLU}'\left(z_t^0\right) = \left(W_y^{\mathrm{T}}\left(\frac{\partial L_t}{\partial y_t} \odot \mathrm{ReLU}'(z_t)\right) + \frac{\partial L_{(t+1)+}}{\partial h_t}\right) \odot \mathrm{ReLU}'\left(z_t^0\right) \tag{16}$$

We can write the derivatives as:

$$\frac{\partial L_t}{\partial W_y} = \delta_t\, h_t^{\mathrm{T}} \qquad\qquad \frac{\partial L_t}{\partial b_y} = \delta_t \tag{17}$$

$$\frac{\partial L_t}{\partial W_x} = \delta_t^h\, x_t^{\mathrm{T}} \qquad\qquad \frac{\partial L_t}{\partial b_h} = \delta_t^h \tag{18}$$

$$\frac{\partial L_t}{\partial W_h} = \delta_t^h\, h_{t-1}^{\mathrm{T}} \tag{19}$$

And the backpropagated errors through time and to the previous layer are:

$$\frac{\partial L_{t+}}{\partial h_{t-1}} = W_h^{\mathrm{T}}\,\delta_t^h \tag{20}$$

$$\frac{\partial L_t}{\partial x_t} = W_x^{\mathrm{T}}\,\delta_t^h \tag{21}$$

When backpropagating through time, (21) plays the role of the second term in (11) in the previous time step. Remark: If we were to calculate the recursive gradients in each time step, we get the following formulae:

$$\frac{\partial L_t}{\partial W_x} = \frac{\partial L_t}{\partial h_t}\sum_{i=1}^{t}\left(W_h^{\mathrm{T}\,t-1}\left(\prod_{j=1}^{i}\mathrm{ReLU}'\left(z_{t-j+1}^0\right)\right) x_{t-i+1}^{\mathrm{T}}\right) \tag{22}$$

$$\frac{\partial L_t}{\partial W_h} = \frac{\partial L_t}{\partial h_t}\sum_{i=1}^{t}\left(W_h^{\mathrm{T}\,t-1}\left(\prod_{j=1}^{i}\mathrm{ReLU}'\left(z_{t-j+1}^0\right)\right) h_{t-i}^{\mathrm{T}}\right) \tag{23}$$

$$\frac{\partial L_t}{\partial b_h} = \frac{\partial L_t}{\partial h_t}\sum_{i=1}^{t}\left(W_h^{\mathrm{T}\,t-1}\left(\prod_{j=1}^{i}\mathrm{ReLU}'\left(z_{t-j+1}^0\right)\right)\right) \tag{24}$$

2. LSTM backpropagation

The LSTM unit is given by the equations:

$$f_t = \sigma\big(W_f x_t + U_f h_{t-1} + b_f\big) \tag{25}$$
$$i_t = \sigma\big(W_i x_t + U_i h_{t-1} + b_i\big) \tag{26}$$
$$o_t = \sigma\big(W_o x_t + U_o h_{t-1} + b_o\big) \tag{27}$$
$$\tilde{c}_t = \tanh\big(W_c x_t + U_c h_{t-1} + b_c\big) \tag{28}$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{29}$$
$$h_t = o_t \odot \tanh\big(c_t\big) \tag{30}$$

The derivatives of the L_t loss terms w.r.t. parameters are:

$$\frac{\partial L_t}{\partial W_c} = \frac{\partial L_t}{\partial h_t}\frac{\partial h_t}{\partial W_c} = \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{\tanh(c_t)\frac{\partial o_t}{\partial W_c} + o_t\frac{\partial \tanh(c_t)}{\partial W_c}\right\} =$$
$$= \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{\tanh(c_t)\left\{o_t\Big(1-o_t\Big)U_o\frac{\partial h_{t-1}}{\partial W_c}\right\} + o_t\left\{\Big(1-\tanh^2(c_t)\Big)\frac{\partial c_t}{\partial W_c}\right\}\right\} =$$
$$= \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{\tanh(c_t)\left\{o_t\Big(1-o_t\Big)U_o\frac{\partial h_{t-1}}{\partial W_c}\right\} + o_t\left\{\Big(1-\tanh^2(c_t)\Big)\left\{c_{t-1}\frac{\partial f_t}{\partial W_c} + f_t\frac{\partial c_{t-1}}{\partial W_c} + \tilde{c}_t\frac{\partial i_t}{\partial W_c} + i_t\frac{\partial \tilde{c}_t}{\partial W_c}\right\}\right\}\right\} =$$

$$\begin{aligned}
&= \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{\tanh(c_t)\left\{o_t\Big(1-o_t\Big)U_o\frac{\partial h_{t-1}}{\partial W_c}\right\}\right\} + \\
&+ \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{o_t\Big(1-\tanh^2(c_t)\Big)c_{t-1}\left\{f_t\Big(1-f_t\Big)U_f\frac{\partial h_{t-1}}{\partial W_c}\right\}\right\} + \\
&+ \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{o_t\Big(1-\tanh^2(c_t)\Big)f_t\frac{\partial c_{t-1}}{\partial W_c}\right\} + \\
&+ \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{o_t\Big(1-\tanh^2(c_t)\Big)\tilde{c}_t\left\{i_t\Big(1-i_t\Big)U_i\frac{\partial h_{t-1}}{\partial W_c}\right\}\right\} + \\
&+ \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{o_t\Big(1-\tanh^2(c_t)\Big)i_t\left\{\Big(1-\tilde{c}_t^2\Big)x_t\right\}\right\} =
\end{aligned}$$

$$\begin{aligned}
&= U_o^{\mathrm{T}}\left(\frac{\partial L_t}{\partial h_t}\odot\tanh(c_t)\odot o_t\odot\Big(1-o_t\Big)\right)\frac{\partial h_{t-1}}{\partial W_c} + \\
&+ U_f^{\mathrm{T}}\left(\frac{\partial L_t}{\partial h_t}\odot o_t\odot\Big(1-\tanh^2(c_t)\Big)\odot c_{t-1}\odot f_t\odot\Big(1-f_t\Big)\right)\frac{\partial h_{t-1}}{\partial W_c} + \\
&+ \left(\frac{\partial L_t}{\partial h_t}\odot o_t\odot\Big(1-\tanh^2(c_t)\Big)\odot f_t\right)\frac{\partial c_{t-1}}{\partial W_c} + \\
&+ U_i^{\mathrm{T}}\left(\frac{\partial L_t}{\partial h_t}\odot o_t\odot\Big(1-\tanh^2(c_t)\Big)\odot\tilde{c}_t\odot i_t\odot\Big(1-i_t\Big)\right)\frac{\partial h_{t-1}}{\partial W_c} + \\
&+ \left(\frac{\partial L_t}{\partial h_t}\odot o_t\odot\Big(1-\tanh^2(c_t)\Big)\odot i_t\odot\Big(1-\tilde{c}_t^2\Big)\right)x_t^{\mathrm{T}}
\end{aligned} \tag{31}$$

$$
\begin{aligned}
&\frac{\partial L_t}{\partial W_o} = \frac{\partial L_t}{\partial h_t}\frac{\partial h_t}{\partial W_o} = \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{\tanh(c_t)\frac{\partial o_t}{\partial W_o} + o_t\frac{\partial \tanh(c_t)}{\partial W_o}\right\} = \\
&= \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{\tanh(c_t)\left\{o_t\Big(1-o_t\Big)\Big(U_o\frac{\partial h_{t-1}}{\partial W_o} + x_t\Big)\right\} + o_t\left\{\Big(1-\tanh^2(c_t)\Big)\frac{\partial c_t}{\partial W_o}\right\}\right\} = \\
&= \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{\tanh(c_t)\left\{o_t\Big(1-o_t\Big)\Big(U_o\frac{\partial h_{t-1}}{\partial W_o} + x_t\Big)\right\} + o_t\left\{\Big(1-\tanh^2(c_t)\Big)\left\{c_{t-1}\frac{\partial f_t}{\partial W_o} + f_t\frac{\partial c_{t-1}}{\partial W_o} + \tilde{c}_t\frac{\partial i_t}{\partial W_o} + i_t\frac{\partial \tilde{c}_t}{\partial W_o}\right\}\right\}\right\} = \\
&= \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{\tanh(c_t)\left\{o_t\Big(1-o_t\Big)x_t^{\mathrm{T}}\right\}\right\} + \\
&+ \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{\tanh(c_t)\left\{o_t\Big(1-o_t\Big)U_o\frac{\partial h_{t-1}}{\partial W_o}\right\}\right\} + \\
&+ \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{o_t\Big(1-\tanh^2(c_t)\Big)c_{t-1}\left\{f_t\Big(1-f_t\Big)U_f\frac{\partial h_{t-1}}{\partial W_o}\right\}\right\} + \\
&+ \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{o_t\Big(1-\tanh^2(c_t)\Big)f_t\frac{\partial c_{t-1}}{\partial W_o}\right\} + \\
&+ \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{o_t\Big(1-\tanh^2(c_t)\Big)\tilde{c}_t\left\{i_t\Big(1-i_t\Big)U_i\frac{\partial h_{t-1}}{\partial W_o}\right\}\right\} + \\
&+ \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{o_t\Big(1-\tanh^2(c_t)\Big)i_t\left\{\Big(1-\tilde{c}_t^2\Big)U_c\frac{\partial h_{t-1}}{\partial W_o}\right\}\right\} = \\
&= \left(\frac{\partial L_t}{\partial h_t}\odot\tanh(c_t)\odot o_t\odot\Big(1-o_t\Big)\right)x_t^{\mathrm{T}} + \\
&+ U_o^{\mathrm{T}}\left(\frac{\partial L_t}{\partial h_t}\odot\tanh(c_t)\odot o_t\odot\Big(1-o_t\Big)\right)\frac{\partial h_{t-1}}{\partial W_o} + \\
&+ U_f^{\mathrm{T}}\left(\frac{\partial L_t}{\partial h_t}\odot o_t\odot\Big(1-\tanh^2(c_t)\Big)\odot c_{t-1}\odot f_t\odot\Big(1-f_t\Big)\right)\frac{\partial h_{t-1}}{\partial W_o} + \\
&+ \left(\frac{\partial L_t}{\partial h_t}\odot o_t\odot\Big(1-\tanh^2(c_t)\Big)\odot f_t\right)\frac{\partial c_{t-1}}{\partial W_o} + \\
&+ U_i^{\mathrm{T}}\left(\frac{\partial L_t}{\partial h_t}\odot o_t\odot\Big(1-\tanh^2(c_t)\Big)\odot \tilde{c}_t\odot i_t\odot\Big(1-i_t\Big)\right)\frac{\partial h_{t-1}}{\partial W_o} + \\
&+ U_c^{\mathrm{T}}\left(\frac{\partial L_t}{\partial h_t}\odot o_t\odot\Big(1-\tanh^2(c_t)\Big)\odot i_t\odot\Big(1-\tilde{c}_t^2\Big)\right)\frac{\partial h_{t-1}}{\partial W_o} \qquad (32)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial L_t}{\partial W_i} &= \frac{\partial L_t}{\partial h_t}\frac{\partial h_t}{\partial W_i} = \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{\tanh(c_t)\frac{\partial o_t}{\partial W_i} + o_t\frac{\partial \tanh(c_t)}{\partial W_i}\right\} = \\
&= \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{\tanh(c_t)\left\{o_t\Big(1-o_t\Big)U_o\frac{\partial h_{t-1}}{\partial W_i}\right\} + o_t\left\{\Big(1-\tanh^2(c_t)\Big)\frac{\partial c_t}{\partial W_i}\right\}\right\} = \\
&= \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{\tanh(c_t)\left\{o_t\Big(1-o_t\Big)U_o\frac{\partial h_{t-1}}{\partial W_i}\right\} + o_t\left\{\Big(1-\tanh^2(c_t)\Big)\left\{c_{t-1}\frac{\partial f_t}{\partial W_i} + f_t\frac{\partial c_{t-1}}{\partial W_i} + \tilde{c}_t\frac{\partial i_t}{\partial W_i} + i_t\frac{\partial \tilde{c}_t}{\partial W_i}\right\}\right\}\right\} =
\end{aligned}
$$

$$
\begin{aligned}
&= \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{\tanh(c_t)\left\{o_t\Big(1-o_t\Big)U_o\frac{\partial h_{t-1}}{\partial W_i}\right\}\right\} + \\
&+ \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{o_t\Big(1-\tanh^2(c_t)\Big)c_{t-1}\left\{f_t\Big(1-f_t\Big)U_f\frac{\partial h_{t-1}}{\partial W_i}\right\}\right\} + \\
&+ \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{o_t\Big(1-\tanh^2(c_t)\Big)f_t\frac{\partial c_{t-1}}{\partial W_i}\right\} + \\
&+ \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{o_t\Big(1-\tanh^2(c_t)\Big)\tilde{c}_t\left\{i_t\Big(1-i_t\Big)\Big(x_t + U_i\frac{\partial h_{t-1}}{\partial W_i}\Big)\right\}\right\} + \\
&+ \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{o_t\Big(1-\tanh^2(c_t)\Big)i_t\left\{\Big(1-\tilde{c}_t^2\Big)U_c\frac{\partial h_{t-1}}{\partial W_i}\right\}\right\} =
\end{aligned}
$$

$$
\begin{aligned}
&= U_o^{\mathrm{T}}\left(\frac{\partial L_t}{\partial h_t}\odot\tanh(c_t)\odot o_t\odot\Big(1-o_t\Big)\right)\frac{\partial h_{t-1}}{\partial W_i} + \\
&+ U_f^{\mathrm{T}}\left(\frac{\partial L_t}{\partial h_t}\odot o_t\odot\Big(1-\tanh^2(c_t)\Big)\odot c_{t-1}\odot f_t\odot\Big(1-f_t\Big)\right)\frac{\partial h_{t-1}}{\partial W_i} + \\
&+ \quad\left(\frac{\partial L_t}{\partial h_t}\odot o_t\odot\Big(1-\tanh^2(c_t)\Big)\odot f_t\right)\frac{\partial c_{t-1}}{\partial W_i} + \\
&+ U_i^{\mathrm{T}}\left(\frac{\partial L_t}{\partial h_t}\odot o_t\odot\Big(1-\tanh^2(c_t)\Big)\odot\tilde{c}_t\odot i_t\odot\Big(1-i_t\Big)\right)\frac{\partial h_{t-1}}{\partial W_i} + \\
&+ \quad\left(\frac{\partial L_t}{\partial h_t}\odot o_t\odot\Big(1-\tanh^2(c_t)\Big)\odot\tilde{c}_t\odot i_t\odot\Big(1-i_t\Big)\right)x_t^{\mathrm{T}} + \\
&+ U_c^{\mathrm{T}}\left(\frac{\partial L_t}{\partial h_t}\odot o_t\odot\Big(1-\tanh^2(c_t)\Big)\odot i_t\odot\Big(1-\tilde{c}_t^2\Big)\right)\frac{\partial h_{t-1}}{\partial W_i}
\end{aligned}
\tag{33}
$$

$$
\begin{aligned}
\frac{\partial L_t}{\partial W_f} = \frac{\partial L_t}{\partial h_t}\frac{\partial h_t}{\partial W_f} = \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{\tanh(c_t)\frac{\partial o_t}{\partial W_f} + o_t\frac{\partial \tanh(c_t)}{\partial W_f}\right\} = \\
= \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{\tanh(c_t)\left\{o_t\Big(1-o_t\Big)U_o\frac{\partial h_{t-1}}{\partial W_f}\right\} + o_t\left\{\Big(1-\tanh^2(c_t)\Big)\frac{\partial c_t}{\partial W_f}\right\}\right\} = \\
= \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{\tanh(c_t)\left\{o_t\Big(1-o_t\Big)U_o\frac{\partial h_{t-1}}{\partial W_f}\right\} + o_t\left\{\Big(1-\tanh^2(c_t)\Big)\left\{c_{t-1}\frac{\partial f_t}{\partial W_f} + f_t\frac{\partial c_{t-1}}{\partial W_f} + \tilde{c}_t\frac{\partial i_t}{\partial W_f} + i_t\frac{\partial \tilde{c}_t}{\partial W_f}\right\}\right\}\right\} =
\end{aligned}
$$

$$
\begin{aligned}
&= \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{\tanh(c_t)\left\{o_t\Big(1-o_t\Big)U_o\frac{\partial h_{t-1}}{\partial W_f}\right\}\right\} + \\
&+ \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{o_t\Big(1-\tanh^2(c_t)\Big)c_{t-1}\left\{f_t\Big(1-f_t\Big)\Big(x_t + U_f\frac{\partial h_{t-1}}{\partial W_f}\Big)\right\}\right\} + \\
&+ \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{o_t\Big(1-\tanh^2(c_t)\Big)f_t\frac{\partial c_{t-1}}{\partial W_f}\right\} + \\
&+ \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{o_t\Big(1-\tanh^2(c_t)\Big)\tilde{c}_t\left\{i_t\Big(1-i_t\Big)U_i\frac{\partial h_{t-1}}{\partial W_f}\right\}\right\} + \\
&+ \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{o_t\Big(1-\tanh^2(c_t)\Big)i_t\left\{\Big(1-\tilde{c}_t^2\Big)U_c\frac{\partial h_{t-1}}{\partial W_f}\right\}\right\} =
\end{aligned}
$$

$$
\begin{aligned}
&= U_o^{\mathrm{T}}\left(\frac{\partial L_t}{\partial h_t}\odot\tanh(c_t)\odot o_t\odot\Big(1-o_t\Big)\right)\frac{\partial h_{t-1}}{\partial W_f} + \\
&+ \quad \left(\frac{\partial L_t}{\partial h_t}\odot o_t\odot\Big(1-\tanh^2(c_t)\Big)\odot\tilde{c}_{t-1}\odot f_t\odot\Big(1-f_t\Big)\right)x_t^{\mathrm{T}} + \\
&+ U_f^{\mathrm{T}}\left(\frac{\partial L_t}{\partial h_t}\odot o_t\odot\Big(1-\tanh^2(c_t)\Big)\odot c_{t-1}\odot f_t\odot\Big(1-f_t\Big)\right)\frac{\partial h_{t-1}}{\partial W_f} + \\
&+ \quad \left(\frac{\partial L_t}{\partial h_t}\odot o_t\odot\Big(1-\tanh^2(c_t)\Big)\odot f_t\right)\frac{\partial c_{t-1}}{\partial W_f} + \\
&+ U_i^{\mathrm{T}}\left(\frac{\partial L_t}{\partial h_t}\odot o_t\odot\Big(1-\tanh^2(c_t)\Big)\odot\tilde{c}_t\odot i_t\odot\Big(1-i_t\Big)\right)\frac{\partial h_{t-1}}{\partial W_f} + \\
&+ U_c^{\mathrm{T}}\left(\frac{\partial L_t}{\partial h_t}\odot o_t\odot\Big(1-\tanh^2(c_t)\Big)\odot i_t\odot\Big(1-\tilde{c}_t^2\Big)\right)\frac{\partial h_{t-1}}{\partial W_f}
\end{aligned}
\tag{34}
$$

The backpropagated errors are:

$$
\begin{aligned}
\frac{\partial L_t}{\partial h_{t-1}} = \frac{\partial L_t}{\partial h_t}\frac{\partial h_t}{\partial h_{t-1}} = \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{\tanh(c_t)\frac{\partial o_t}{\partial h_{t-1}} + o_t\frac{\partial \tanh(c_t)}{\partial h_{t-1}}\right\} = \\
= \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{\tanh(c_t)\left\{o_t\Big(1-o_t\Big)U_o\right\} + o_t\left\{\Big(1-\tanh^2(c_t)\Big)\frac{\partial c_t}{\partial h_{t-1}}\right\}\right\} = \\
= \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{\tanh(c_t)\left\{o_t\Big(1-o_t\Big)U_o\right\} + o_t\left\{\Big(1-\tanh^2(c_t)\Big)\left\{c_{t-1}\frac{\partial f_t}{\partial h_{t-1}} + \tilde{c}_t\frac{\partial i_t}{\partial h_{t-1}} + i_t\frac{\partial \tilde{c}_t}{\partial h_{t-1}}\right\}\right\}\right\} =
\end{aligned}
$$

$$
\begin{aligned}
&= \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{\tanh(c_t)\left\{o_t\Big(1-o_t\Big)U_o\right\}\right\} + \\
&+ \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{o_t\Big(1-\tanh^2(c_t)\Big)c_{t-1}\left\{f_t\Big(1-f_t\Big)\Big(x_t+U_f\Big)\right\}\right\} + \\
&+ \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{o_t\Big(1-\tanh^2(c_t)\Big)\tilde{c}_t\left\{i_t\Big(1-i_t\Big)U_i\right\}\right\} + \\
&+ \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{o_t\Big(1-\tanh^2(c_t)\Big)i_t\left\{\Big(1-\tilde{c}_t^2\Big)U_c\right\}\right\} = \\
\\
&= U_o^{\mathrm{T}}\left(\frac{\partial L_t}{\partial h_t}\odot\tanh(c_t)\odot o_t\odot\Big(1-o_t\Big)\right) + \\
&+ U_f^{\mathrm{T}}\left(\frac{\partial L_t}{\partial h_t}\odot o_t\odot\Big(1-\tanh^2(c_t)\Big)\odot c_{t-1}\odot f_t\odot\Big(1-f_t\Big)\right) + \\
&+ U_i^{\mathrm{T}}\left(\frac{\partial L_t}{\partial h_t}\odot o_t\odot\Big(1-\tanh^2(c_t)\Big)\odot\tilde{c}_t\odot i_t\odot\Big(1-i_t\Big)\right) + \\
&+ U_c^{\mathrm{T}}\left(\frac{\partial L_t}{\partial h_t}\odot o_t\odot\Big(1-\tanh^2(c_t)\Big)\odot i_t\odot\Big(1-\tilde{c}_t^2\Big)\right)
\end{aligned}
\tag{35}
$$

$$
\begin{aligned}
\frac{\partial L_t}{\partial c_{t-1}} &= \frac{\partial L_t}{\partial h_t}\frac{\partial h_t}{\partial c_{t-1}} = \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{\tanh(c_t)\frac{\partial o_t}{\partial c_{t-1}} + o_t\frac{\partial\tanh(c_t)}{\partial c_{t-1}}\right\} = \\
&= \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{\tanh(c_t)\left\{o_t\Big(1-o_t\Big)U_o\frac{\partial h_{t-1}}{\partial c_{t-1}}\right\} + o_t\left\{\Big(1-\tanh^2(c_t)\Big)\frac{\partial c_t}{\partial c_{t-1}}\right\}\right\} = \\
&= \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{\tanh(c_t)\left\{o_t\Big(1-o_t\Big)U_o\frac{\partial h_{t-1}}{\partial c_{t-1}}\right\} + o_t\left\{\Big(1-\tanh^2(c_t)\Big)\left\{c_{t-1}\frac{\partial f_t}{\partial c_{t-1}} + f_t + \tilde{c}_t\frac{\partial i_t}{\partial c_{t-1}} + i_t\frac{\partial\tilde{c}_t}{\partial c_{t-1}}\right\}\right\}\right\} = \\
\\
&= \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{\tanh(c_t)\left\{o_t\Big(1-o_t\Big)U_o\frac{\partial h_{t-1}}{\partial c_{t-1}}\right\}\right\} + \\
&+ \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{o_t\Big(1-\tanh^2(c_t)\Big)c_{t-1}\left\{f_t\Big(1-f_t\Big)U_f\frac{\partial h_{t-1}}{\partial c_{t-1}}\right\}\right\} + \\
&+ \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{o_t\Big(1-\tanh^2(c_t)\Big)f_t\right\} + \\
&+ \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{o_t\Big(1-\tanh^2(c_t)\Big)\tilde{c}_t\left\{i_t\Big(1-i_t\Big)U_i\frac{\partial h_{t-1}}{\partial c_{t-1}}\right\}\right\} + \\
&+ \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{o_t\Big(1-\tanh^2(c_t)\Big)i_t\left\{\Big(1-\tilde{c}_t^2\Big)U_c\frac{\partial h_{t-1}}{\partial c_{t-1}}\right\}\right\} = \\
\\
&= U_o^{\mathrm{T}}\left(\frac{\partial L_t}{\partial h_t}\odot\tanh(c_t)\odot o_t\odot\Big(1-o_t\Big)\right) + \\
&+ U_f^{\mathrm{T}}\left(\frac{\partial L_t}{\partial h_t}\odot o_t\odot\Big(1-\tanh^2(c_t)\Big)\odot c_{t-1}\odot f_t\odot\Big(1-f_t\Big)\right)\frac{\partial h_{t-1}}{\partial c_{t-1}} + \\
&+ \quad\left(\frac{\partial L_t}{\partial h_t}\odot o_t\odot\Big(1-\tanh^2(c_t)\Big)\odot f_t\right) + \\
&+ U_i^{\mathrm{T}}\left(\frac{\partial L_t}{\partial h_t}\odot o_t\odot\Big(1-\tanh^2(c_t)\Big)\odot\tilde{c}_t\odot i_t\odot\Big(1-i_t\Big)\right)\frac{\partial h_{t-1}}{\partial c_{t-1}} + \\
&+ U_c^{\mathrm{T}}\left(\frac{\partial L_t}{\partial h_t}\odot o_t\odot\Big(1-\tanh^2(c_t)\Big)\odot i_t\odot\Big(1-\tilde{c}_t^2\Big)\right)\frac{\partial h_{t-1}}{\partial c_{t-1}}
\end{aligned}
\tag{36}
$$

$$
\frac{\partial L_t}{\partial x_t} = \frac{\partial L_t}{\partial h_t}\frac{\partial h_t}{\partial x_t} = \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{\tanh(c_t)\frac{\partial o_t}{\partial x_t} + o_t\frac{\partial \tanh(c_t)}{\partial x_t}\right\} =
$$

$$
= \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{\tanh(c_t)\left\{o_t\Big(1-o_t\Big)W_o\right\} + o_t\left\{\Big(1-\tanh^2(c_t)\Big)\frac{\partial c_t}{\partial x_t}\right\}\right\} =
$$

$$
= \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{\tanh(c_t)\left\{o_t\Big(1-o_t\Big)W_o\right\} + o_t\left\{\Big(1-\tanh^2(c_t)\Big)\left\{c_{t-1}\frac{\partial f_t}{\partial x_t} + \tilde{c}_t\frac{\partial i_t}{\partial x_t} + i_t\frac{\partial \tilde{c}_t}{\partial x_t}\right\}\right\}\right\} =
$$

$$
\begin{aligned}
&= \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{\tanh(c_t)\left\{o_t\Big(1-o_t\Big)W_o\right\}\right\} + \\
&+ \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{o_t\Big(1-\tanh^2(c_t)\Big)c_{t-1}\left\{f_t\Big(1-f_t\Big)W_f\right\}\right\} + \\
&+ \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{o_t\Big(1-\tanh^2(c_t)\Big)\tilde{c}_t\left\{i_t\Big(1-i_t\Big)W_i\right\}\right\} + \\
&+ \left\{\frac{\partial L_t}{\partial h_t}\right\}\left\{o_t\Big(1-\tanh^2(c_t)\Big)i_t\left\{\Big(1-\tilde{c}_t^2\Big)W_c\right\}\right\} =
\end{aligned}
$$

$$
\begin{aligned}
&= W_o^{\mathrm{T}}\left(\frac{\partial L_t}{\partial h_t} \odot \tanh(c_t) \odot o_t \odot \Big(1-o_t\Big)\right) + \\
&+ W_f^{\mathrm{T}}\left(\frac{\partial L_t}{\partial h_t} \odot o_t \odot \Big(1-\tanh^2(c_t)\Big) \odot c_{t-1} \odot f_t \odot \Big(1-f_t\Big)\right) + \\
&+ W_i^{\mathrm{T}}\left(\frac{\partial L_t}{\partial h_t} \odot o_t \odot \Big(1-\tanh^2(c_t)\Big) \odot \tilde{c}_t \odot i_t \odot \Big(1-i_t\Big)\right) + \\
&+ W_c^{\mathrm{T}}\left(\frac{\partial L_t}{\partial h_t} \odot o_t \odot \Big(1-\tanh^2(c_t)\Big) \odot i_t \odot \Big(1-\tilde{c}_t^2\Big)\right) \qquad (37)
\end{aligned}
$$

By introducing the variables:

$$
\delta_t^o = \frac{\partial L_t}{\partial h_t} \odot \tanh(c_t) \odot o_t \odot \Big(1-o_t\Big) \qquad (38)
$$

$$
\delta_t^f = \frac{\partial L_t}{\partial h_t} \odot o_t \odot \Big(1-\tanh^2(c_t)\Big) \odot c_{t-1} \odot f_t \odot \Big(1-f_t\Big) \qquad (39)
$$

$$
\delta_t^i = \frac{\partial L_t}{\partial h_t} \odot o_t \odot \Big(1-\tanh^2(c_t)\Big) \odot \tilde{c}_t \odot i_t \odot \Big(1-i_t\Big) \qquad (40)
$$

$$
\delta_t^c = \frac{\partial L_t}{\partial h_t} \odot o_t \odot \Big(1-\tanh^2(c_t)\Big) \odot f_t \qquad (41)
$$

$$
\delta_t^{\tilde{c}} = \frac{\partial L_t}{\partial h_t} \odot o_t \odot \Big(1-\tanh^2(c_t)\Big) \odot i_t \odot \Big(1-\tilde{c}_t^2\Big) \qquad (42)
$$

The derivatives w.r.t. the parameters become:

$$
\frac{\partial L_t}{\partial W_c} = \delta_t^{\tilde{c}} x_t^{\mathrm{T}} + \Big(U_o^{\mathrm{T}}\delta_t^o + U_f^{\mathrm{T}}\delta_t^f + U_i^{\mathrm{T}}\delta_t^i\Big)\frac{\partial h_{t-1}}{\partial W_c} + \delta_t^c\frac{\partial c_{t-1}}{\partial W_c} \qquad (43)
$$

$$
\frac{\partial L_t}{\partial U_c} = \delta_t^{\tilde{c}} h_{t-1}^{\mathrm{T}} + \Big(U_o^{\mathrm{T}}\delta_t^o + U_f^{\mathrm{T}}\delta_t^f + U_i^{\mathrm{T}}\delta_t^i\Big)\frac{\partial h_{t-1}}{\partial U_c} + \delta_t^c\frac{\partial c_{t-1}}{\partial U_c} \qquad (44)
$$

$$
\frac{\partial L_t}{\partial b_c} = \delta_t^{\tilde{c}} + \Big(U_o^{\mathrm{T}}\delta_t^o + U_f^{\mathrm{T}}\delta_t^f + U_i^{\mathrm{T}}\delta_t^i\Big)\frac{\partial h_{t-1}}{\partial b_c} + \delta_t^c\frac{\partial c_{t-1}}{\partial b_c} \qquad (45)
$$

$$\frac{\partial L_t}{\partial W_o} = \delta_t^o x_t^{\mathrm{T}} + \left(U_o^{\mathrm{T}}\delta_t^o + U_f^{\mathrm{T}}\delta_t^f + U_i^{\mathrm{T}}\delta_t^i + U_c^{\mathrm{T}}\delta_t^{\tilde{c}}\right)\frac{\partial h_{t-1}}{\partial W_o} + \delta_t^c\frac{\partial c_{t-1}}{\partial W_o} \tag{46}$$

$$\frac{\partial L_t}{\partial U_o} = \delta_t^o h_{t-1}^{\mathrm{T}} + \left(U_o^{\mathrm{T}}\delta_t^o + U_f^{\mathrm{T}}\delta_t^f + U_i^{\mathrm{T}}\delta_t^i + U_c^{\mathrm{T}}\delta_t^{\tilde{c}}\right)\frac{\partial h_{t-1}}{\partial U_o} + \delta_t^c\frac{\partial c_{t-1}}{\partial U_o} \tag{47}$$

$$\frac{\partial L_t}{\partial b_o} = \delta_t^o + \left(U_o^{\mathrm{T}}\delta_t^o + U_f^{\mathrm{T}}\delta_t^f + U_i^{\mathrm{T}}\delta_t^i + U_c^{\mathrm{T}}\delta_t^{\tilde{c}}\right)\frac{\partial h_{t-1}}{\partial b_o} + \delta_t^c\frac{\partial c_{t-1}}{\partial b_o} \tag{48}$$

$$\frac{\partial L_t}{\partial W_i} = \delta_t^i x_t^{\mathrm{T}} + \left(U_o^{\mathrm{T}}\delta_t^o + U_f^{\mathrm{T}}\delta_t^f + U_i^{\mathrm{T}}\delta_t^i + U_c^{\mathrm{T}}\delta_t^{\tilde{c}}\right)\frac{\partial h_{t-1}}{\partial W_i} + \delta_t^c\frac{\partial c_{t-1}}{\partial W_i} \tag{49}$$

$$\frac{\partial L_t}{\partial U_i} = \delta_t^i h_{t-1}^{\mathrm{T}} + \left(U_o^{\mathrm{T}}\delta_t^o + U_f^{\mathrm{T}}\delta_t^f + U_i^{\mathrm{T}}\delta_t^i + U_c^{\mathrm{T}}\delta_t^{\tilde{c}}\right)\frac{\partial h_{t-1}}{\partial U_i} + \delta_t^c\frac{\partial c_{t-1}}{\partial U_i} \tag{50}$$

$$\frac{\partial L_t}{\partial b_i} = \delta_t^i + \left(U_o^{\mathrm{T}}\delta_t^o + U_f^{\mathrm{T}}\delta_t^f + U_i^{\mathrm{T}}\delta_t^i + U_c^{\mathrm{T}}\delta_t^{\tilde{c}}\right)\frac{\partial h_{t-1}}{\partial b_i} + \delta_t^c\frac{\partial c_{t-1}}{\partial b_i} \tag{51}$$

$$\frac{\partial L_t}{\partial W_f} = \delta_t^f x_t^{\mathrm{T}} + \left(U_o^{\mathrm{T}}\delta_t^o + U_f^{\mathrm{T}}\delta_t^f + U_i^{\mathrm{T}}\delta_t^i + U_c^{\mathrm{T}}\delta_t^{\tilde{c}}\right)\frac{\partial h_{t-1}}{\partial W_f} + \delta_t^c\frac{\partial c_{t-1}}{\partial W_f} \tag{52}$$

$$\frac{\partial L_t}{\partial U_f} = \delta_t^f h_{t-1}^{\mathrm{T}} + \left(U_o^{\mathrm{T}}\delta_t^o + U_f^{\mathrm{T}}\delta_t^f + U_i^{\mathrm{T}}\delta_t^i + U_c^{\mathrm{T}}\delta_t^{\tilde{c}}\right)\frac{\partial h_{t-1}}{\partial U_f} + \delta_t^c\frac{\partial c_{t-1}}{\partial U_f} \tag{53}$$

$$\frac{\partial L_t}{\partial b_f} = \delta_t^f x_t^{\mathrm{T}} + \left(U_o^{\mathrm{T}}\delta_t^o + U_f^{\mathrm{T}}\delta_t^f + U_i^{\mathrm{T}}\delta_t^i + U_c^{\mathrm{T}}\delta_t^{\tilde{c}}\right)\frac{\partial h_{t-1}}{\partial b_f} + \delta_t^c\frac{\partial c_{t-1}}{\partial b_f} \tag{54}$$

$$\frac{\partial L_t}{\partial h_{t-1}} = \left(U_o^{\mathrm{T}}\delta_t^o + U_f^{\mathrm{T}}\delta_t^f + U_i^{\mathrm{T}}\delta_t^i + U_c^{\mathrm{T}}\delta_t^{\tilde{c}}\right) \tag{55}$$

$$\frac{\partial L_t}{\partial c_{t-1}} = \delta_t^c + \left(U_o^{\mathrm{T}}\delta_t^o + U_f^{\mathrm{T}}\delta_t^f + U_i^{\mathrm{T}}\delta_t^i + U_c^{\mathrm{T}}\delta_t^{\tilde{c}}\right)\frac{\partial h_{t-1}}{\partial c_{t-1}} = \delta_t^c + \frac{\partial L_t}{\partial h_{t-1}}\frac{\partial h_{t-1}}{\partial c_{t-1}} \tag{56}$$

$$\frac{\partial L_t}{\partial x_t} = \left(W_o^{\mathrm{T}}\delta_t^o + W_f^{\mathrm{T}}\delta_t^f + W_i^{\mathrm{T}}\delta_t^i + W_c^{\mathrm{T}}\delta_t^{\tilde{c}}\right) \tag{57}$$

3. GRU backpropagation

The gated recurrent unit is given by the equations:

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \tag{58}$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \tag{59}$$

$$\tilde{h}_t = \tanh(W_h x_t + U_h (r_t \odot h_{t-1}) + b_h) \tag{60}$$

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \tilde{h}_t \tag{61}$$

By the same token as in the RNN case, we can account for the contribution of the future time steps by adding $\frac{\partial L_{(t+1)+}}{\partial h_t}$ to the first term. This way, we only need to consider the functions belonging to the current time step. The derivatives of the L_t loss terms w.r.t. parameters become:

$$\frac{\partial L_{t+}}{\partial h_t} = \frac{\partial L_t}{\partial h_t} + \frac{\partial L_{(t+1)+}}{\partial h_t} \tag{62}$$

$$\begin{aligned}\frac{\partial L_t}{\partial W_h} &= \frac{\partial L_{t+}}{\partial h_t}\frac{\partial h_t}{\partial W_h} = \frac{\partial L_{t+}}{\partial h_t}\frac{\partial h_t}{\partial \tilde{h}_t}\frac{\partial \tilde{h}_t}{\partial W_h} = \left\{\frac{\partial L_{t+}}{\partial h_t}\right\}\left\{1 - z_t\right\}\left\{\left(1 - \tilde{h}_t^2\right)\, x_t\right\} = \\ &= \frac{\partial L_{t+}}{\partial h_t} \odot \left(1 - z_t\right) \odot \left(1 - \tilde{h}_t^2\right)\, x_t^{\mathrm{T}}\end{aligned} \tag{63}$$

$$\begin{aligned}\frac{\partial L_t}{\partial U_h} &= \frac{\partial L_{t+}}{\partial h_t}\frac{\partial h_t}{\partial U_h} = \frac{\partial L_{t+}}{\partial h_t}\frac{\partial h_t}{\partial \tilde{h}_t}\frac{\partial \tilde{h}_t}{\partial U_h} = \left\{\frac{\partial L_{t+}}{\partial h_t}\right\}\left\{1 - z_t\right\}\left\{\left(1 - \tilde{h}_t^2\right)\left(r_t\, h_{t-1}\right)\right\} = \\ &= \frac{\partial L_{t+}}{\partial h_t} \odot \left(1 - z_t\right) \odot \left(1 - \tilde{h}_t^2\right)\left(r_t \odot h_{t-1}\right)^{\mathrm{T}}\end{aligned} \tag{64}$$

$$\begin{aligned}\frac{\partial L_t}{\partial b_h} &= \frac{\partial L_{t+}}{\partial h_t}\frac{\partial h_t}{\partial b_h} = \frac{\partial L_{t+}}{\partial h_t}\frac{\partial h_t}{\partial \tilde{h}_t}\frac{\partial \tilde{h}_t}{\partial b_h} = \left\{\frac{\partial L_{t+}}{\partial h_t}\right\}\left\{1 - z_t\right\}\left\{1 - \tilde{h}_t^2\right\} = \\ &= \frac{\partial L_{t+}}{\partial h_t} \odot \left(1 - z_t\right) \odot \left(1 - \tilde{h}_t^2\right)\end{aligned} \tag{65}$$

$$\begin{aligned}\frac{\partial L_{t+}}{\partial W_r} &= \frac{\partial L_{t+}}{\partial h_t}\frac{\partial h_t}{\partial W_r} = \frac{\partial L_{t+}}{\partial h_t}\frac{\partial h_t}{\partial \tilde{h}_t}\frac{\partial \tilde{h}_t}{\partial r_t}\frac{\partial r_t}{\partial W_r} = \left\{\frac{\partial L_{t+}}{\partial h_t}\right\}\left\{1 - z_t\right\}\left\{\left(1 - \tilde{h}_t^2\right) U_h\, h_{t-1}\right\}\left\{r_t \left(1 - r_t\right) x_t\right\} = \\ &= U_h^{\mathrm{T}} \frac{\partial L_{t+}}{\partial h_t} \odot \left(1 - z_t\right) \odot \left(1 - \tilde{h}_t^2\right) \odot h_{t-1} \odot r_t \odot \left(1 - r_t\right) x_t^{\mathrm{T}}\end{aligned} \tag{66}$$

$$\begin{aligned}\frac{\partial L_{t+}}{\partial U_r} &= \frac{\partial L_{t+}}{\partial h_t}\frac{\partial h_t}{\partial U_r} = \frac{\partial L_{t+}}{\partial h_t}\frac{\partial h_t}{\partial \tilde{h}_t}\frac{\partial \tilde{h}_t}{\partial r_t}\frac{\partial r_t}{\partial U_r} = \left\{\frac{\partial L_{t+}}{\partial h_t}\right\}\left\{1 - z_t\right\}\left\{\left(1 - \tilde{h}_t^2\right) U_h\, h_{t-1}\right\}\left\{r_t \left(1 - r_t\right) h_{t-1}\right\} = \\ &= U_h^{\mathrm{T}} \frac{\partial L_{t+}}{\partial h_t} \odot \left(1 - z_t\right) \odot \left(1 - \tilde{h}_t^2\right) \odot h_{t-1} \odot r_t \odot \left(1 - r_t\right) h_{t-1}^{\mathrm{T}}\end{aligned} \tag{67}$$

$$\begin{aligned}\frac{\partial L_{t+}}{\partial b_r} &= \frac{\partial L_{t+}}{\partial h_t}\frac{\partial h_t}{\partial b_r} = \frac{\partial L_{t+}}{\partial h_t}\frac{\partial h_t}{\partial \tilde{h}_t}\frac{\partial \tilde{h}_t}{\partial r_t}\frac{\partial r_t}{\partial b_r} = \left\{\frac{\partial L_{t+}}{\partial h_t}\right\}\left\{1 - z_t\right\}\left\{\left(1 - \tilde{h}_t^2\right) U_h\, h_{t-1}\right\}\left\{r_t \left(1 - r_t\right)\right\} = \\ &= U_h^{\mathrm{T}} \frac{\partial L_{t+}}{\partial h_t} \odot \left(1 - z_t\right) \odot \left(1 - \tilde{h}_t^2\right) \odot h_{t-1} \odot r_t \odot \left(1 - r_t\right)\end{aligned} \tag{68}$$

$$\frac{\partial L_{t+}}{\partial W_z} = \frac{\partial L_{t+}}{\partial h_t}\frac{\partial h_t}{\partial W_z} = \frac{\partial L_{t+}}{\partial h_t}\frac{\partial h_t}{\partial z_t}\frac{\partial z_t}{\partial W_z} = \left\{\frac{\partial L_{t+}}{\partial h_t}\right\}\left\{h_{t-1} - \tilde{h}_t\right\}\left\{z_t\left(1 - z_t\right)x_t\right\} =$$
$$= U_h^{\mathrm{T}}\frac{\partial L_{t+}}{\partial h_t} \odot \left(1 - z_t\right) \odot \left(1 - \tilde{h}_t^2\right) \odot h_{t-1} \odot r_t \odot \left(1 - r_t\right) x_t^{\mathrm{T}} \tag{69}$$

$$\frac{\partial L_{t+}}{\partial U_z} = \frac{\partial L_{t+}}{\partial h_t}\frac{\partial h_t}{\partial U_z} = \frac{\partial L_{t+}}{\partial h_t}\frac{\partial h_t}{\partial z_t}\frac{\partial z_t}{\partial U_z} = \left\{\frac{\partial L_{t+}}{\partial h_t}\right\}\left\{h_{t-1} - \tilde{h}_t\right\}\left\{z_t\left(1 - z_t\right)h_{t-1}\right\} =$$
$$= U_h^{\mathrm{T}}\frac{\partial L_{t+}}{\partial h_t} \odot \left(1 - z_t\right) \odot \left(1 - \tilde{h}_t^2\right) \odot h_{t-1} \odot r_t \odot \left(1 - r_t\right) h_{t-1}^{\mathrm{T}} \tag{70}$$

$$\frac{\partial L_{t+}}{\partial U_z} = \frac{\partial L_{t+}}{\partial h_t}\frac{\partial h_t}{\partial U_z} = \frac{\partial L_{t+}}{\partial h_t}\frac{\partial h_t}{\partial z_t}\frac{\partial z_t}{\partial U_z} = \left\{\frac{\partial L_{t+}}{\partial h_t}\right\}\left\{h_{t-1} - \tilde{h}_t\right\}\left\{z_t\left(1 - z_t\right)h_{t-1}\right\} =$$
$$= U_h^{\mathrm{T}}\frac{\partial L_{t+}}{\partial h_t} \odot \left(1 - z_t\right) \odot \left(1 - \tilde{h}_t^2\right) \odot h_{t-1} \odot r_t \odot \left(1 - r_t\right) \tag{71}$$

And the backpropagated errors are:

$$\frac{\partial L_{t+}}{\partial h_{t-1}} = \frac{\partial L_{t+}}{\partial h_t}\frac{\partial h_t}{\partial h_{t-1}} = \left\{\frac{\partial L_{t+}}{\partial h_t}\right\}\left\{z_t + \frac{\partial z_t}{\partial h_{t-1}}h_{t-1} - \frac{\partial z_t}{\partial h_{t-1}}\tilde{h}_t + \left(1 - z_t\right)\frac{\partial \tilde{h}_t}{\partial h_{t-1}}\right\} =$$
$$= \left\{\frac{\partial L_{t+}}{\partial h_t}\right\}\left\{z_t + \left(h_{t-1} - \tilde{h}_t\right)\frac{\partial z_t}{\partial h_{t-1}} + \left(1 - z_t\right)\frac{\partial \tilde{h}_t}{\partial h_{t-1}}\right\} =$$
$$= \left\{\frac{\partial L_{t+}}{\partial h_t}\right\}\left\{z_t + \left(h_{t-1} - \tilde{h}_t\right)\left\{z_T\left(1 - z_t\right)U_z\right\} + \left(1 - z_t\right)\left\{\left(1 - \tilde{h}_t^2\right)U_h\left(\frac{\partial r_t}{\partial h_{t-1}}h_{t-1} + r_t\right)\right\}\right\} =$$
$$= \left\{\frac{\partial L_{t+}}{\partial h_t}\right\}\left\{z_t + \left(h_{t-1} - \tilde{h}_t\right)\left\{z_T\left(1 - z_t\right)U_z\right\} + \left(1 - z_t\right)\left\{\left(1 - \tilde{h}_t^2\right)U_h\left(\left\{r_t\left(1 - r_t\right)U_r\right\}h_{t-1} + r_t\right)\right\}\right\} =$$
$$= \frac{\partial L_{t+}}{\partial h_t} \odot z_t + U_z^{\mathrm{T}}\frac{\partial L_{t+}}{\partial h_t} \odot \left(h_{t-1} - \tilde{h}_t\right) \odot z_t \odot \left(1 - z_t\right) + U_h^{\mathrm{T}}\frac{\partial L_{t+}}{\partial h_t} \odot \left(1 - z_t\right) \odot \left(1 - \tilde{h}_t^2\right) \odot r_t +$$
$$+ U_r^{\mathrm{T}}U_h^{\mathrm{T}}\frac{\partial L_{t+}}{\partial h_t} \odot \left(1 - z_t\right) \odot \left(1 - \tilde{h}_t^2\right) \odot h_{t-1} \odot r_t \odot \left(1 - r_t\right) \tag{72}$$

$$\frac{\partial L_{t+}}{\partial x_t} = \frac{\partial L_{t+}}{\partial h_t}\frac{\partial h_t}{\partial x_t} = \left\{\frac{\partial L_{t+}}{\partial h_t}\right\}\left\{\frac{\partial z_t}{\partial x_t}h_{t-1} - \frac{\partial z_t}{\partial x_t}\tilde{h}_t + \frac{\partial \tilde{h}_t}{\partial x_t}\left(1 - z_t\right)\right\} = \left\{\frac{\partial L_{t+}}{\partial h_t}\right\}\left\{\frac{\partial z_t}{\partial x_t}\left(h_{t-1} - \tilde{h}_t\right) + \frac{\partial \tilde{h}_t}{\partial x_t}\left(1 - z_t\right)\right\} =$$
$$= \left\{\frac{\partial L_{t+}}{\partial h_t}\right\}\left\{\left\{z_t\left(1 - z_t\right)W_z\right\}\left(h_{t-1} - \tilde{h}_t\right) + \left\{\left(1 - \tilde{h}_t^2\right)\left(W_h + U_h\tilde{h}_t\frac{\partial r_t}{\partial x_t}\right)\right\}\left(1 - z_t\right)\right\} =$$
$$= \left\{\frac{\partial L_{t+}}{\partial h_t}\right\}\left\{\left\{z_t\left(1 - z_t\right)W_z\right\}\left(h_{t-1} - \tilde{h}_t\right) + \left\{\left(1 - \tilde{h}_t^2\right)\left(W_h + U_h\tilde{h}_t\left\{r_t\left(1 - r_t\right)W_r\right\}\right)\right\}\left(1 - z_t\right)\right\} =$$
$$= W_z^{\mathrm{T}}\frac{\partial L_{t+}}{\partial h_t} \odot \left(h_{t-1} - \tilde{h}_t\right) \odot z_t \odot \left(1 - z_t\right) + W_h^{\mathrm{T}}\frac{\partial L_{t+}}{\partial h_t} \odot \left(1 - z_t\right) \odot \left(1 - \tilde{h}_t^2\right) +$$
$$+ W_r^{\mathrm{T}}U_h^{\mathrm{T}}\frac{\partial L_{t+}}{\partial h_t} \odot \left(1 - z_t\right) \odot \left(1 - \tilde{h}_t^2\right) \odot h_{t-1} \odot r_t \odot \left(1 - r_t\right) \tag{73}$$

By introducing the notations:

$$\delta_t^z = \frac{\partial L_{t+}}{\partial h_t} \odot \left(h_{t-1} - \tilde{h}_t\right) \odot z_t \odot \left(1 - z_t\right) \tag{74}$$
$$\delta_t^h = \frac{\partial L_{t+}}{\partial h_t} \odot \left(1 - z_t\right) \odot \left(1 - \tilde{h}_t^2\right) \tag{75}$$
$$\delta_t^r = U_h^{\mathrm{T}}\frac{\partial L_{t+}}{\partial h_t} \odot \left(1 - z_t\right) \odot \left(1 - \tilde{h}_t^2\right) \odot h_{t-1} \odot r_t \odot \left(1 - r_t\right) \tag{76}$$

The gradients can be written as:

$$\frac{\partial L_{t+}}{\partial W_h} = \delta_t^h \, x_t^{\mathrm{T}} \tag{77}$$

$$\frac{\partial L_{t+}}{\partial U_h} = \delta_t^h \left(r_t \odot h_{t-1}\right)^{\mathrm{T}} \tag{78}$$

$$\frac{\partial L_{t+}}{\partial b_h} = \delta_t^h \tag{79}$$

$$\frac{\partial L_{t+}}{\partial W_r} = \delta_t^r \, x_t^{\mathrm{T}} \tag{80}$$

$$\frac{\partial L_{t+}}{\partial U_r} = \delta_t^r \, h_{t-1}^{\mathrm{T}} \tag{81}$$

$$\frac{\partial L_{t+}}{\partial b_r} = \delta_t^r \tag{82}$$

$$\frac{\partial L_{t+}}{\partial W_z} = \delta_t^z \, x_t^{\mathrm{T}} \tag{83}$$

$$\frac{\partial L_{t+}}{\partial U_z} = \delta_t^z \, h_{t-1}^{\mathrm{T}} \tag{84}$$

$$\frac{\partial L_{t+}}{\partial U_z} = \delta_t^z \tag{85}$$

$$\frac{\partial L_{t+}}{\partial h_{t-1}} = \frac{\partial L_{t+}}{\partial h_t} \odot z_t + U_z^{\mathrm{T}} \, \delta_t^z + U_h^{\mathrm{T}} \, \delta_t^z \odot r_t + U_r^{\mathrm{T}} \delta_t^r \tag{86}$$

$$\frac{\partial L_{t+}}{\partial x_t} = W_z^{\mathrm{T}} \, \delta_t^z + W_h^{\mathrm{T}} \, \delta_t^h + W_r^{\mathrm{T}} U_h^{\mathrm{T}} \, \delta_t^r \tag{87}$$

$\frac{\partial L_{t+}}{\partial x_t}$ goes into (31) in the previous time step and we utilized the sigmoid and tanh derivatives.