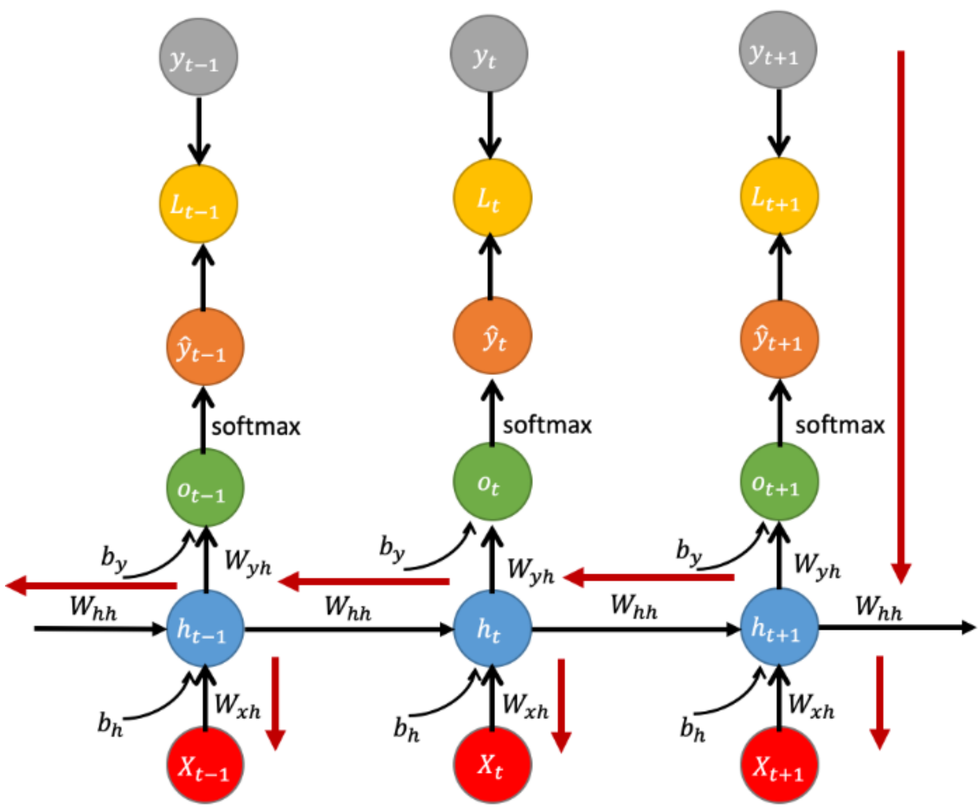
Backpropagation through time for RNN

Qazi Zarif Ul Islam

March 25, 2024



Unfolded Recurrent Neural Net
(Credit: [2])

In this document, we shall derive gradient of the loss of RNN w.r.t. the *hidden weights* W_{hh}. In truth, we shall derive the gradient aka differentiate the **loss at a particular time instant wrt *the concatenated weight matrix*** $[W_{xh}\ W_{hh}]^T$ This concatenation does not make any difference from a computational perspective. [1, 2].
The emprical loss of the neural network is,

$$\hat{L} = \frac{1}{T}\sum_{t=1}^{T}\ell(y_t, d_t) = \frac{1}{T}(l_1 + l_2 + ... + l_t ... + l_T) \tag{1}$$

For an RNN, the system is defined by,

$$h_t = f(X_t, h_{t-1}) = \phi_h(W_{xh} \cdot X_t + W_{hh} \cdot h_{t-1} + b_h) \tag{2}$$

$$\hat{o}_t = f_o(h_t) = \phi_o(W_{hy} \cdot h_t + b_y) \tag{3}$$

Let $w_h = [W_{xh}\ W_{hh}]^T$.
Thus,

$$\frac{\partial \hat{L}}{\partial w_h} = \frac{1}{T}(\frac{\partial l_1}{\partial w_h} + \frac{\partial l_2}{\partial w_h} + ... + \frac{\partial l_t}{\partial w_h} ... + \frac{\partial l_T}{\partial w_h}) \tag{4}$$

Now the loss at a particular time instant t follows the chain rule of derivatives.

$$\frac{\partial l_t}{\partial w_h} = \frac{\partial l_t}{\partial o_t}\frac{\partial o_t}{\partial h_t}\frac{\partial h_t}{\partial w_h} \tag{5}$$

But h_t is a function of h_{t-1} and w_h as well (besides X_t). Furthermore, h_{t-1} is again a function of w_h. Thus, by the multivariable chain rule,
If $h_t = f_1(h_{t-1}, w_h), h_{t-1} = g_1(h_{t-2}, w_h), w_h = h_1(w_h)$ [1]

$$\frac{\partial h_t}{\partial w_h} = \frac{\partial f_1}{\partial w_h} + \frac{\partial f_1}{\partial h_{t-1}}\frac{\partial h_{t-1}}{\partial w_h} \tag{6}$$

Similarly, if $h_{t-1} = f_2(h_{t-2}, w_h), h_{t-2} = g_2(h_{t-3}, w_h), w_h = h_2(w_h)$

$$\frac{\partial h_{t-1}}{\partial w_h} = \frac{\partial f_2}{\partial w_h} + \frac{\partial f_2}{\partial h_{t-2}}\frac{\partial h_{t-2}}{\partial w_h} \tag{7}$$

Similarly, if $h_{t-2} = f_3(h_{t-3}, w_h), h_{t-3} = g_3(h_{t-4}, w_h), w_h = h_3(w_h)$

$$\frac{\partial h_{t-2}}{\partial w_h} = \frac{\partial f_3}{\partial w_h} + \frac{\partial f_3}{\partial h_{t-3}}\frac{\partial h_{t-3}}{\partial w_h} \tag{8}$$

We can find similar expressions for time steps further back h_{t-3}, h_{t-4} all the way to h_1.
Thus, eqn 6 becomes,

[1](Note: h_1 (the function, *not* the hidden state) is merely the identity function. We've formulated it so just to be consistent with the formulation of "case 1" in [3]. Eqn 6 would still be valid even if we hadn't defined it as a separate function.

$$\frac{\partial h_t}{\partial w_h} = \frac{\partial f_1}{\partial w_h} + \frac{\partial f_1}{\partial h_{t-1}}(\frac{\partial f_2}{\partial w_h} + \frac{\partial f_2}{\partial h_{t-2}}\frac{\partial h_{t-2}}{\partial w_h}) \qquad \text{(we won't expand } \frac{\partial h_{t-2}}{\partial w_h}) \tag{9}$$

$$= \frac{\partial f_1}{\partial w_h} + \frac{\partial f_1}{\partial h_{t-1}}\frac{\partial f_2}{\partial w_h} + \frac{\partial f_1}{\partial h_{t-1}}\frac{\partial f_2}{\partial h_{t-2}}\frac{\partial h_{t-2}}{\partial w_h} \tag{10}$$

$$= \frac{\partial h_t}{\partial w_h} + \frac{\partial h_t}{\partial h_{t-1}}\frac{\partial h_{t-1}}{\partial w_h} + \frac{\partial h_t}{\partial h_{t-1}}\frac{\partial h_{t-1}}{\partial h_{t-2}}\frac{\partial h_{t-2}}{\partial w_h} \tag{11}$$

Now, we need to expand $\frac{\partial h_{t-2}}{\partial w_h}$ using eqn 8 to get the final form but we can instead write it more compactly by realizing that this equation is ultimately a **summation of multiplications**. Let's break it down next.

What are the terms of the summation? Observe that every term's last (right-most) term is partial of h_i and i progresses from the beginning and ends at t as we move from right to left in the summation. Hence, i is our index of summation. Now, every term is again a cascade of multiplications, decreasing in number of multipliers as we move from right to left in the summation. Every multiplier term is a partial of h_j wrt h_{j-1}. (If our index of multiplication was i, as it was for the summation, each summation term would be different from those found in eqn 6.)

Thus, the final form of 6 can be written as,

$$\boxed{\frac{\partial h_t}{\partial w_h} = \frac{\partial h_t}{\partial w_h} + \sum_{i=1}^{t-1}\prod_{j-1=i}^{t}(\frac{\partial h_j}{\partial h_{j-1}})\frac{\partial h_i}{\partial w_h}}$$

Which is the same as equation 9.7.7 in [1].

References

[1] *Backpropagation Through Time*. `https://d2l.ai/chapter_recurrent-neural-networks/bptt.html`. Accessed: 2024-03-24.

[2] Murat Karakaya. *Backpropagation Through Time of RNN*. `https://mmuratarat.github.io/2019-02-07/bptt-of-rnn`. Accessed: 2024-03-24. 2019.

[3] Paul's online notes. *Chain rule*. `https://tutorial.math.lamar.edu/classes/calciii/chainrule.aspx`. Accessed: 2024-03-24. 2018.