1 Recurrent Neural Network

A recurrent neural network (RNN) is a class of artificial neural network where connections between units form a directed cycle. This creates an internal state of the network which allows it to exhibit dynamic temporal behaviour. The type of network that we will be interested in this document resembles:

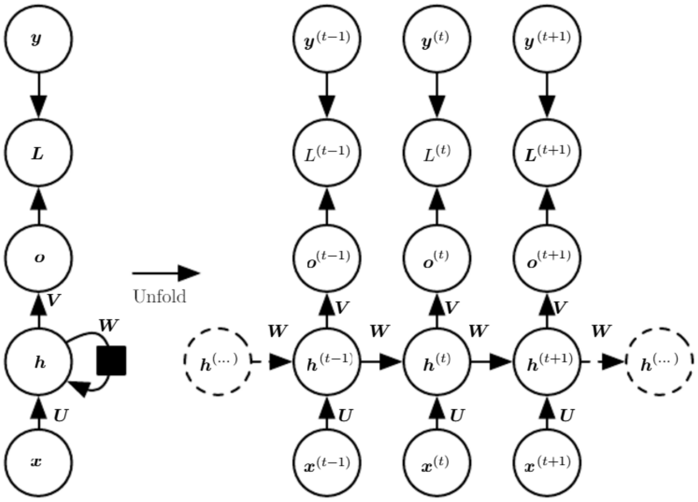

Figure 1: Recurrent Neural Network (RNN)

At a fixed time step a recurrent neural network behaves similarly to a feed forward network, where the hidden layer is connected from one time step to the next. That is for a given time step we have:

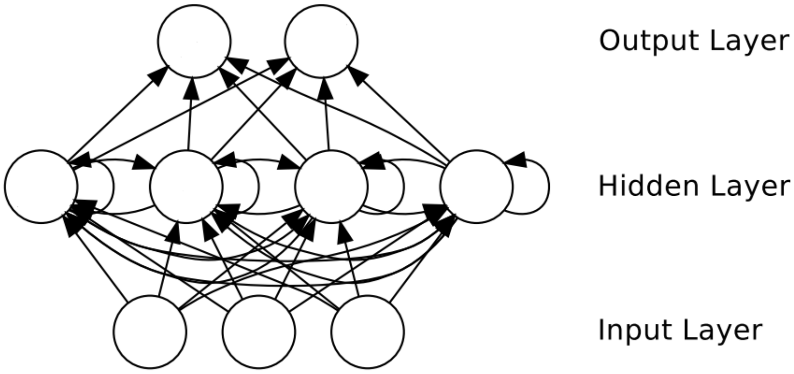

Figure 2: RNN at a given time step [2]

For a simple recurrent network with one hidden node, we will characterize our network as follows:

$$\mathbf{a}^{(t)} = \mathbf{b} + \mathbf{W}\mathbf{h}^{(t-1)} + \mathbf{U}\mathbf{x}^{(t)} \tag{1}$$

$$\mathbf{h}^{(t)} = \tanh(\mathbf{a}^{(t)}) \tag{2}$$

$$\mathbf{o}^{(t)} = \mathbf{c} + \mathbf{V}\mathbf{h}^{(t)} \tag{3}$$

$$\hat{\mathbf{y}}^{(t)} = \mathrm{softmax}(\mathbf{o}^{(t)}) \tag{4}$$

To help with the derivation of the feed forward procedure and the backpropagation algorithm we will let the inputs $\mathbf{x}^{(t)} \in \mathbb{R}^4$, the outputs $\mathbf{o}^{(t)} \in \mathbb{R}^3$ and the actual values $\mathbf{y}^{(t)} \in \mathbb{R}^3$, furthermore we let the weights $\mathbf{U} \in \mathbb{R}^{5\times4}, \mathbf{W} \in \mathbb{R}^{5\times5}$ and $\mathbf{V} \in \mathbb{R}^{3\times5}$, and finally the biases $\mathbf{b} \in \mathbb{R}^5$, and $\mathbf{c} \in \mathbb{R}^3$. This allows us to write the activation $\mathbf{a}^{(t)}$ as:

$$\begin{pmatrix} a_1^{(t)} \\ a_2^{(t)} \\ a_3^{(t)} \\ a_4^{(t)} \\ a_5^{(t)} \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{pmatrix} + \begin{pmatrix} w_{11} & w_{12} & w_{13} & w_{14} & w_{15} \\ w_{21} & w_{22} & w_{23} & w_{24} & w_{25} \\ w_{31} & w_{32} & w_{33} & w_{34} & w_{35} \\ w_{41} & w_{42} & w_{43} & w_{44} & w_{45} \\ w_{51} & w_{52} & w_{53} & w_{54} & w_{55} \end{pmatrix} \begin{pmatrix} h_1^{(t-1)} \\ h_2^{(t-1)} \\ h_3^{(t-1)} \\ h_4^{(t-1)} \\ h_5^{(t-1)} \end{pmatrix} + \begin{pmatrix} u_{11} & u_{12} & u_{13} & u_{14} \\ u_{21} & u_{22} & u_{23} & u_{24} \\ u_{31} & u_{32} & u_{33} & u_{34} \\ u_{41} & u_{42} & u_{43} & u_{44} \\ u_{51} & u_{52} & u_{53} & u_{54} \end{pmatrix} \begin{pmatrix} x_1^{(t)} \\ x_2^{(t)} \\ x_3^{(t)} \\ x_4^{(t)} \end{pmatrix} \tag{5}$$

we can be write the above a little more compactly as

$$\begin{aligned} a_1^{(t)} &= b_1 + \sum_{i=1}^{5} w_{1i} h_i^{(t-1)} + \sum_{j=1}^{4} u_{1j} x_j^{(t)} \qquad (6) \\ a_2^{(t)} &= b_2 + \sum_{i=1}^{5} w_{2i} h_i^{(t-1)} + \sum_{j=1}^{4} u_{2j} x_j^{(t)} \\ a_3^{(t)} &= b_3 + \sum_{i=1}^{5} w_{3i} h_i^{(t-1)} + \sum_{j=1}^{4} u_{3j} x_j^{(t)} \\ a_4^{(t)} &= b_4 + \sum_{i=1}^{5} w_{4i} h_i^{(t-1)} + \sum_{j=1}^{4} u_{4j} x_j^{(t)} \\ a_5^{(t)} &= b_5 + \sum_{i=1}^{5} w_{5i} h_i^{(t-1)} + \sum_{j=1}^{4} u_{5j} x_j^{(t)} \end{aligned}$$

We could do the same for the output, so that we arrive at

$$\begin{pmatrix} o_1^{(t)} \\ o_2^{(t)} \\ o_3^{(t)} \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} + \begin{pmatrix} v_{11} & v_{12} & v_{13} & v_{14} & v_{15} \\ v_{21} & v_{22} & v_{23} & v_{24} & v_{25} \\ v_{31} & v_{32} & v_{33} & v_{34} & v_{35} \end{pmatrix} \begin{pmatrix} \tanh\left(a_1^{(t)}\right) \\ \tanh\left(a_2^{(t)}\right) \\ \tanh\left(a_3^{(t)}\right) \\ \tanh\left(a_4^{(t)}\right) \\ \tanh\left(a_5^{(t)}\right) \end{pmatrix} \tag{7}$$

Which can also be rewritten as:

$$\begin{aligned} o_1^{(t)} &= c_1 + \sum_{i=1}^{5} v_{1i} h_i^{(t)} \qquad (8) \\ o_2^{(t)} &= c_2 + \sum_{i=1}^{5} v_{2i} h_i^{(t)} \\ o_3^{(t)} &= c_3 + \sum_{i=1}^{5} v_{3i} h_i^{(t)} \end{aligned}$$

and the softmax outputs:

$$\begin{pmatrix} \hat{y}_1^{(t)} \\ \hat{y}_2^{(t)} \\ \hat{y}_3^{(t)} \end{pmatrix} = \frac{1}{\sum\limits_{i=1}^{3} \exp\left(o_i^{(t)}\right)} \begin{pmatrix} \exp\left(o_1^{(t)}\right) \\ \exp\left(o_2^{(t)}\right) \\ \exp\left(o_3^{(t)}\right) \end{pmatrix} \tag{9}$$

2 Back-Propagation Through Time

In the derivation of the back-propagation through time algorithm we assume that the outputs $\mathbf{o}^{(t)}$ are used as the argument to the softmax function to obtain the vector $\hat{\mathbf{y}}$ of the probabilities over the output. It is also assumed that the loss is the *negative log-likelihood* of the true target $y^{(t)}$ given the input so far. The total loss for a given sequence of $\mathbf{x}$ values paired with a sequence $\mathbf{y}$ values would then be just the sum of the losses over all the time steps So that [1] [3].

$$\begin{aligned}
&L\big(\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(\tau)}\}, \{\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \dots, \mathbf{y}^{(\tau)}\}\big) && (10)\\
&= \sum_t L^{(t)} && (11)\\
&= -\sum_t \log p_{\text{model}}\big(y^{(t)} \big| \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(t)}\}\big) && (12)
\end{aligned}$$

In the case of our example the loss at a given time period is expressed as:

$$L^{(t)} = -\sum_{i=1}^{3} y_i^{(t)} \ln\left(\hat{y}_i^{(t)}\right) \tag{13}$$

expanded out:

$$\begin{aligned}
&= -\left(y_1^{(t)} \ln\left(\hat{y}_1^{(t)}\right) + y_2^{(t)} \ln\left(\hat{y}_2^{(t)}\right) + y_3^{(t)} \ln\left(\hat{y}_3^{(t)}\right)\right) && (14)\\
&= -\left[y_1^{(t)} \ln\left(\frac{\exp\left(o_1^{(t)}\right)}{\sum_{i=1}^{3} \exp\left(o_i^{(t)}\right)}\right) + y_2^{(t)} \ln\left(\frac{\exp\left(o_2^{(t)}\right)}{\sum_{i=1}^{3} \exp\left(o_i^{(t)}\right)}\right) + y_3^{(t)} \ln\left(\frac{\exp\left(o_3^{(t)}\right)}{\sum_{i=1}^{3} \exp\left(o_i^{(t)}\right)}\right)\right] && (15)\\
&= -\Bigg[y_1^{(t)}\left(o_1^{(t)} - \ln\left(\sum_{i=1}^{3} \exp\left(o_i^{(t)}\right)\right)\right) + y_2^{(t)}\left(o_2^{(t)} - \ln\left(\sum_{i=1}^{3} \exp\left(o_i^{(t)}\right)\right)\right)\\
&\qquad + y_3^{(t)}\left(o_3^{(t)} - \ln\left(\sum_{i=1}^{3} \exp\left(o_i^{(t)}\right)\right)\right)\Bigg] && (16)
\end{aligned}$$

When computing the gradient $\nabla_{\mathbf{o}^{(t)}} L$ on the outputs at time step t for all i, t we notice from figure 1 that perturbing $\mathbf{o}^{(t)}$ will change $L^{(t)}$ at a given time and have no effect on the loss of $L^{(t)}$ for $s \neq t$. The gradient of L with respect to $\mathbf{o}^{(t)}$ will depend on the loss function, in our case we are using the *negative log-likelihood*. Without loss of generality, let us assume that the activated class for the given time is 1, so that we arrive at:

$$L^{(t)} = -\ln\left(\hat{y}_1^{(t)}\right) = o_1^{(t)} - \ln\left(\sum_{i=1}^{3} \exp\left(o_i^{(t)}\right)\right) \tag{17}$$

From (11) we see that:

$$\frac{\partial L}{\partial L^{(t)}} = 1 \tag{18}$$

then taking $\nabla_{\mathbf{o}^{(t)}} L$ we arrive at

$$\begin{aligned}
\nabla_{\mathbf{o}^{(t)}} L &= -\left(\left(1 - \frac{\exp(o_1^{(t)})}{\sum_{i=1}^{3} \exp\left(o_i^{(t)}\right)}\right) \quad -\frac{\exp(o_2^{(t)})}{\sum_{i=1}^{3} \exp\left(o_i^{(t)}\right)} \quad -\frac{\exp(o_3^{(t)})}{\sum_{i=1}^{3} \exp\left(o_i^{(t)}\right)}\right)^T && (19)\\
&= -\left(\left(1 - \hat{y}_1^{(t)}\right) \quad -\hat{y}_2^{(t)} \quad -\hat{y}_3^{(t)}\right)^T && (20)
\end{aligned}$$

$$= \left(\left(\hat{y}_1^{(t)} - 1\right) \quad \hat{y}_2^{(t)} \quad \hat{y}_3^{(t)}\right)^T \tag{21}$$

For a class k we can write this to be calculated to be:

$$\left(\nabla_{\mathbf{o}^{(t)}} L\right)_i = \frac{\partial L}{\partial o_i^{(t)}} = \frac{\partial L}{\partial L^{(t)}} \frac{L^{(t)}}{\partial o_i^{(t)}} = \hat{y}_i^{(t)} - \mathbb{1}_{i,y^{(t)}} \tag{22}$$

Where $\mathbb{1}_{\text{condition}} = 1$ if the condition is true, else is zero.

We take a look at computing the total loss in terms of the activations $\mathbf{h}^{(t)}$. We notice that perturbing $\mathbf{h}^{(t)}$ has the effect of changing $\mathbf{o}^{(t)}$ which in turn changes $L^{(t)}$ and $\mathbf{h}^{(t)}$ propagates to $\mathbf{h}^{(t+1)}, \mathbf{h}^{(t+2)}, \dots \mathbf{h}^{(\tau)}$, changing the loss at all times greater then t. We begin by computing $\nabla_{\mathbf{h}^{(\tau)}} L$ at the final time step τ, since it only has $o^{(t)}$ as a descendent. [1]

We start computing from (17) and write

$$L^{(\tau)} = -\ln\left(\hat{y}_1^{(\tau)}\right) = -o_1^{(\tau)} + \ln\left(\sum_{i=1}^{3} \exp\left(o_i^{(\tau)}\right)\right) \tag{23}$$

$$= -c_1 - \sum_{i=1}^{5} v_{1i} h_i^{(\tau)} + \ln\left(\sum_{k=1}^{3} \exp\left(c_k + \sum_{l=1}^{5} v_{kl} h_l^{(\tau)}\right)\right) \tag{24}$$

We will take $\frac{\partial L}{\partial h_1^{(\tau)}}$ *(the computation for* $\frac{\partial L}{\partial h_i^{(\tau)}}$ *for* $i = 1, \dots, 5$ *is identical. Combining all partial derivatives will result in our gradient matrix).*

$$\frac{\partial L}{\partial h_1^{(\tau)}} = -v_{11} + \frac{v_{11} \exp\left(c_1 + \sum_{i=1}^{5} v_{1i} h_i^{(\tau)}\right) + v_{21} \exp\left(c_1 + \sum_{i=1}^{5} v_{1i} h_i^{(\tau)}\right) + v_{31} \exp\left(c_1 + \sum_{i=1}^{5} v_{1i} h_i^{(\tau)}\right)}{\sum_{k=1}^{3} \exp\left(c_k + \sum_{l=1}^{5} v_{kl} h_l^{(\tau)}\right)} \tag{25}$$

$$= -v_{11} + v_{11}\hat{y}_1^{(\tau)} + v_{21}\hat{y}_2^{(\tau)} + v_{31}\hat{y}_3^{(\tau)} \tag{26}$$

$$= v_{11}(\hat{y}_1^{(\tau)} - 1) + v_{21}\hat{y}_2^{(\tau)} + v_{31}\hat{y}_3^{(\tau)} \tag{27}$$

repeating the same for $\frac{\partial L}{\partial h_i^{(\tau)}}, i = 2, \dots, 5$ we arrive at:

$$\frac{\partial L}{\partial h_2^{(\tau)}} = v_{12}(\hat{y}_1^{(\tau)} - 1) + v_{22}\hat{y}_2^{(\tau)} + v_{32}\hat{y}_3^{(\tau)} \tag{28}$$

$$\frac{\partial L}{\partial h_3^{(\tau)}} = v_{13}(\hat{y}_1^{(\tau)} - 1) + v_{23}\hat{y}_2^{(\tau)} + v_{31}\hat{y}_3^{(\tau)} \tag{29}$$

$$\frac{\partial L}{\partial h_4^{(\tau)}} = v_{14}(\hat{y}_1^{(\tau)} - 1) + v_{24}\hat{y}_2^{(\tau)} + v_{34}\hat{y}_3^{(\tau)} \tag{30}$$

$$\frac{\partial L}{\partial h_5^{(\tau)}} = v_{15}(\hat{y}_1^{(\tau)} - 1) + v_{25}\hat{y}_2^{(\tau)} + v_{35}\hat{y}_3^{(\tau)} \tag{31}$$

then putting it in matrix for we arrive at:

$$\underbrace{\begin{pmatrix} v_{11} & v_{21} & v_{31} \\ v_{12} & v_{22} & v_{32} \\ v_{13} & v_{23} & v_{33} \\ v_{14} & v_{24} & v_{34} \\ v_{15} & v_{25} & v_{35} \end{pmatrix}}_{\mathbf{V}^T} \underbrace{\begin{pmatrix} \hat{y}_1^{(t)} - 1 \\ \hat{y}_2^{(t)} \\ \hat{y}_3^{(t)} \end{pmatrix}}_{\nabla_{\mathbf{o}}^{(\tau)} L} \tag{32}$$

so that:

$$\nabla_{\mathbf{h}^{(\tau)}} L = \mathbf{V}^T \nabla_{\mathbf{o}^{(\tau)}} L \tag{33}$$

Note that $\mathbf{V}$ *does not depend on the time period that we are in. Furthermore, we are supposed to get a matrix since we are taking the gradient of a vector with respect to another vector.*

We are now ready to back-propagate the gradients through time, from $\tau - 1$ down to $t = 1$. We notice that perturbing $\mathbf{h}^{(t)}$ has the effect of changing $\mathbf{o}^{(t)}$ which in turn changes $L^{(t)}$ and $\mathbf{h}^{(t)}$ propagates to $\mathbf{h}^{(t+1)}, \mathbf{h}^{(t+2)}, \ldots \mathbf{h}^{(\tau)}$, changing the loss at all times greater then t. Since we found the gradient for the last time step in (33) let us now find $\nabla_{\mathbf{h}^{(\mathbf{t})}} L$ for $t = \tau - 1$ which has as descendent both $\mathbf{o}^{(t)}$ and $\mathbf{h}^{(t+1)}$ (in our case $\mathbf{h}^{(t+1)} = \mathbf{h}^{(\tau)}$ - we are considering the second to last time step).

$$\nabla_{\mathbf{h}^{(\tau-1)}} L = \nabla_{\mathbf{h}^{(\tau-1)}} L^{(\tau-1)} + \nabla_{\mathbf{h}^{(\tau-1)}} L^{(\tau)} \tag{34}$$

$$= \underbrace{\nabla_{\mathbf{h}^{(\tau-1)}} L^{(\tau-1)}}_{\mathbf{V}^T \nabla_{\mathbf{o}^{(\tau-1)}} L} + \nabla_{\mathbf{h}^{(\tau-1)}} L^{(\tau)} \tag{35}$$

The first term in (35) is the same as (33) so that we are simply required to compute $\nabla_{\mathbf{h}^{(\tau-1)}} L^{(\tau)}$. We begin our computation from (17) and proceed similarly to the way we computed $\frac{\partial L^{(\tau)}}{\partial h_1^{(\tau)}}$ in (23)

$$L^{(\tau)} = -c_1 - \sum_{i=1}^{5} v_{1i} h_i^{(\tau)} + \ln\left(\sum_{k=1}^{3} \exp\left(c_k + \sum_{l=1}^{5} v_{kl} h_l^{(\tau)}\right)\right) \tag{36}$$

$$= -c_1 - \sum_{i=1}^{5} v_{1i} \tanh\left(b_i + \sum_{\alpha=1}^{5} w_{i\alpha} h_\alpha^{(\tau-1)} + \sum_{\epsilon=1}^{4} u_{i\epsilon} x_\epsilon^{(\tau)})\right) \tag{37}$$

$$+ \ln\left[\sum_{k=1}^{3} \exp\left(c_k + \sum_{l=1}^{5} v_{kl} \tanh\left(b_l + \sum_{\beta=1}^{5} w_{l\beta} h_\beta^{(\tau-1)} + \sum_{\gamma=1}^{4} u_{l\gamma} x_\gamma^{(\tau)})\right)\right)\right]$$

As a reminder:

$$\frac{d}{dx} \tanh x = 1 - \tanh^2(x) \tag{38}$$

Note: due to the long expression the computation will be completed in several part.

$$\frac{\partial}{\partial h_1^{(\tau-1)}} \sum_{i=1}^{5} v_{1i} \tanh\left(b_i + \sum_{\alpha=1}^{5} w_{i\alpha} h_\alpha^{(\tau-1)} + F(\mathbf{x}^{(\tau)})\right) \tag{39}$$

$$= \sum_{i=1}^{5} v_{1i} w_{i1} \left(1 - \underbrace{\tanh^2\left(b_i + \sum_{\alpha=1}^{5} w_{i\alpha} h_\alpha^{(\tau-1)} + F(\mathbf{x}^{(\tau)})\right)}_{(h_1^{(\tau)})^2}\right) \tag{40}$$

$$= \sum_{i=1}^{5} v_{1i} w_{i1} (1 - (h_i^{(\tau)})^2) \tag{41}$$

For the second term we have:

$$\frac{\partial \ln(\cdot)}{\partial h_1^{(t)}} = \frac{\sum_{j=1}^{3} \left[\exp\left(c_j + \cdot\right) \sum_{i=1}^{5} v_{ji} w_{i1} \times \left(1 - (h_i^{(\tau)})^2\right)\right]}{\sum_{k=1}^{3} \exp\left(c_k + \sum_{l=1}^{5} v_{kl} \tanh\left(b_l + \sum_{\beta=1}^{5} w_{l\beta} h_\beta^{(\tau-1)} + F(\mathbf{x}^{(\tau)})\right)\right)} \tag{42}$$

putting it all together with (37) we arrive at

$$\frac{\partial L^{(\tau)}}{\partial h_1^{(\tau-1)}} = \sum_{i=1}^{5}\sum_{j=1}^{3}\left(1-\left(h_i^{(\tau)}\right)^2\right)w_{i1}v_{ji}(\hat{y}_j^{(\tau)}-\mathbb{1}_{1,j}) \tag{43}$$

repeating this for $\frac{\partial L^{(\tau)}}{\partial h_i^{(\tau-1)}}, i = 2, \ldots 5$ and considering the second to last time step we arrive at the following matrix solution:

$$\begin{gathered}\underbrace{\begin{pmatrix} 1-(h_1^{(\tau)})^2 & 0 & 0 & 0 & 0 \\ 0 & 1-(h_2^{(\tau)})^2 & 0 & 0 & 0 \\ 0 & 0 & 1-(h_3^{(\tau)})^2 & 0 & 0 \\ 0 & 0 & 0 & 1-(h_4^{(\tau)})^2 & 0 \\ 0 & 0 & 0 & 0 & 1-(h_5^{(\tau)})^2 \end{pmatrix}}_{\operatorname{diag}\left(1-\left(\mathbf{h}^{(\tau)}\right)^2\right)} \\ \underbrace{\begin{pmatrix} w_{11} & w_{21} & w_{31} & w_{41} & w_{51} \\ w_{12} & w_{22} & w_{32} & w_{42} & w_{52} \\ w_{13} & w_{23} & w_{33} & w_{43} & w_{53} \\ w_{14} & w_{24} & w_{34} & w_{44} & w_{54} \\ w_{15} & w_{25} & w_{35} & w_{45} & w_{55} \end{pmatrix}}_{\mathbf{W}^T} \underbrace{\underbrace{\begin{pmatrix} v_{11} & v_{21} & v_{31} \\ v_{12} & v_{22} & v_{32} \\ v_{13} & v_{23} & v_{33} \\ v_{14} & v_{24} & v_{34} \\ v_{15} & v_{25} & v_{35} \end{pmatrix}}_{\mathbf{V}^T} \underbrace{\begin{pmatrix} \hat{y}_1^{(\tau)}-1 \\ \hat{y}_2^{(\tau)} \\ \hat{y}_3^{(\tau)} \end{pmatrix}}_{\nabla_{\mathbf{o}}^{(\tau)} L}}_{\nabla_{\mathbf{h}^{(\tau)}} L}\end{gathered} \tag{44}$$

We can now back-propagate through time starting from the final time step and working our way back using the following expression:

$$\nabla_{\mathbf{h}^{(t)}} L = \mathbf{W}^T \operatorname{diag}\left(1-\left(\mathbf{h}^{(t+1)}\right)^2\right)\left(\nabla_{\mathbf{h}^{(t+1)}} L\right) + \mathbf{V}^T\left(\nabla_{\mathbf{o}^{(t)}} L\right) \tag{45}$$

Now that we have computed the gradients on the internal nodes we can obtain the gradients on the parameter nodes. Because the parameters are shared across many time times, we must take some care when denoting calculus operations involving these variables. [1]

We begin by finding the gradients with respect to the biases. We start by finding $\nabla_{\mathbf{c}} L$ and return to (17) so that

$$L^{(t)} = -c_1 - \sum_{i=1}^{5} v_{1i}h_i^{(t)} + \ln\left(\sum_{k=1}^{3}\exp\left(c_k + \sum_{l=1}^{5} v_{kl}h_l^{(t)}\right)\right) \tag{46}$$

$$\tag{47}$$

then

$$\frac{\partial L^{(t)}}{\partial c_1} = \hat{y}_1^{(t)} - 1, \qquad \frac{\partial L^{(t)}}{\partial c_2^{(t)}} = \hat{y}_2^{(t)}, \qquad \frac{\partial L^{(t)}}{\partial c_3} = \hat{y}_3^{(t)} \tag{48}$$

and

$$\nabla_{\mathbf{c}^{(t)}} L^{(t)} = \begin{pmatrix} (\hat{y}_1^{(t)}-1) & \hat{y}_2^{(t)} & \hat{y}_3^{(t)} \end{pmatrix}^T = \nabla_{\mathbf{o}^{(t)}} L^{(t)} \tag{49}$$

Hence

$$\nabla_{\mathbf{c}} L = \sum_{t}^{\tau}\left(\frac{\partial \mathbf{o}^{(t)}}{\partial \mathbf{c}}\right)^T \nabla_{\mathbf{o}^{(t)}} L = \sum_{t}^{\tau} \nabla_{\mathbf{o}^{(t)}} L \tag{50}$$

We move to computing $\nabla_{\mathbf{b}} L$, we proceed in a similar fashion to how we computed $\nabla_{\mathbf{h}^{(\mathbf{t})}} L$. We remind the reader that we are assuming that our activation is represented by $(1,0,0)^T$

$$\begin{aligned} L^{(t)} = -c_1 - &\sum_{i=1}^{5} v_{1i} \tanh\left(b_i^{(t)} + \sum_{\alpha=1}^{5} w_{i\alpha} h_{\alpha}^{(t-1)} + F(\mathbf{x}^{(t)})\right) \\ &+ \ln\left[\sum_{k=1}^{3} \exp\left(c_k + \sum_{l=1}^{5} v_{kl} \tanh\left(b_l + \sum_{\beta=1}^{5} w_{l\beta} h_{\beta}^{(t-1)} + F(\mathbf{x}^{(t)})\right)\right)\right] \end{aligned} \tag{51}$$

which for simplicity we write:

$$\begin{aligned} L^{(t)} = -c_1 - &\sum_{i=1}^{5} v_{1i} \tanh\left(b_i + F(\mathbf{h}^{(t-1)},\ \mathbf{x}^{(t)})\right) \\ &+ \ln\left[\sum_{k=1}^{3} \exp\left(c_k + \sum_{l=1}^{5} v_{kl} \tanh\left(b_l + G(\mathbf{h}^{(t-1)},\ \mathbf{x}^{(t)})\right)\right)\right] \end{aligned} \tag{52}$$

The computation is identical to finding $\nabla_{\mathbf{h}^{(t)}} L$ and is performed in (33). We simply need to consider the derivative of $\tanh(\cdot)$ when computing the gradient. At the final time step τ, $\nabla_{\mathbf{b}^{(\tau)}} L$ is

$$\frac{\partial L}{\partial h_1^{(\tau)}} = \left(1 - (h_1^{(\tau)})^2\right)\left(v_{11}(\hat{y}_1^{(\tau)} - 1) + v_{21}\hat{y}_2^{(t)} + v_{32}\hat{y}_3^{(t)}\right) \tag{53}$$

$$\frac{\partial L}{\partial h_2^{(\tau)}} = \left(1 - (h_2^{(\tau)})^2\right)\left(v_{12}(\hat{y}_1^{(\tau)} - 1) + v_{22}\hat{y}_2^{(t)} + v_{32}\hat{y}_3^{(t)}\right) \tag{54}$$

$$\frac{\partial L}{\partial h_3^{(\tau)}} = \left(1 - (h_3^{(\tau)})^2\right)\left(v_{13}(\hat{y}_1^{(\tau)} - 1) + v_{23}\hat{y}_2^{(t)} + v_{31}\hat{y}_3^{(t)}\right) \tag{55}$$

$$\frac{\partial L}{\partial h_4^{(\tau)}} = \left(1 - (h_4^{(\tau)})^2\right)\left(v_{14}(\hat{y}_1^{(\tau)} - 1) + v_{24}\hat{y}_2^{(t)} + v_{34}\hat{y}_3^{(t)}\right) \tag{56}$$

$$\frac{\partial L}{\partial h_5^{(\tau)}} = \left(1 - (h_5^{(\tau)})^2\right)\left(v_{15}(\hat{y}_1^{(\tau)} - 1) + v_{25}\hat{y}_2^{(t)} + v_{35}\hat{y}_3^{(t)}\right) \tag{57}$$

which we can write as

$$\underbrace{\begin{pmatrix} 1-(h_1^{(\tau)})^2 & 0 & 0 & 0 & 0 \\ 0 & 1-(h_2^{(\tau)})^2 & 0 & 0 & 0 \\ 0 & 0 & 1-(h_3^{(\tau)})^2 & 0 & 0 \\ 0 & 0 & 0 & 1-(h_4^{(\tau)})^2 & 0 \\ 0 & 0 & 0 & 0 & 1-(h_5^{(\tau)})^2 \end{pmatrix}}_{\mathrm{diag}\left(1-\left(\mathbf{h}^{(\tau)}\right)^2\right)} \underbrace{\begin{pmatrix} v_{11} & v_{21} & v_{31} \\ v_{12} & v_{22} & v_{32} \\ v_{13} & v_{23} & v_{33} \\ v_{14} & v_{24} & v_{34} \\ v_{15} & v_{25} & v_{35} \end{pmatrix}}_{\mathbf{V}^T} \underbrace{\begin{pmatrix} \hat{y}_1^{(\tau)} - 1 \\ \hat{y}_2^{(\tau)} \\ \hat{y}_3^{(\tau)} \end{pmatrix}}_{\nabla_{\mathbf{o}}^{(\tau)} L} \tag{58}$$

so that

$$\begin{aligned} \nabla_{\mathbf{b}} L^{(\tau)} &= \mathrm{diag}\left(1 - \left(\mathbf{h}^{(\tau)}\right)^2\right) \mathbf{V}^T \nabla_{\mathbf{o}^{(\tau)}} L && (59) \\ &= \mathrm{diag}\left(1 - \left(\mathbf{h}^{(\tau)}\right)^2\right) \nabla_{\mathbf{h}^{(\tau)}} L && (60) \end{aligned}$$

The computation is not shown but for the time steps $t = \tau - 1$ we have:

$$\nabla_{\mathbf{b}} L^{(\tau-1)} = \mathrm{diag}\left(1 - (\mathbf{h}^{(\tau-1)})^2\right) \underbrace{\left(\mathbf{W}^T\left(\nabla_{\mathbf{h}^{(\tau)}} L\right)\mathrm{diag}\left(1 - (\mathbf{h}^{(\tau)})^2\right) + \mathbf{V}^T\left(\nabla_{\mathbf{o}^{(\tau)}} L\right)\right)}_{\nabla_{\mathbf{h}^{(\tau-1)}} L} \tag{61}$$

$$= \operatorname{diag}\Big(1 - \big(\mathbf{h}^{(\tau-1)}\big)^2\Big)\nabla_{\mathbf{h}^{(\tau-1)}} L \tag{62}$$

for an arbitrary time step the $\nabla_{\mathbf{b}} L$

$$\nabla_{\mathbf{b}} L = \sum_{t}^{\tau} \left(\frac{\partial \mathbf{h}^{(t)}}{\partial \mathbf{b}}\right)^T \nabla_{\mathbf{h}^{(t)}} L = \sum_{t}^{\tau} \operatorname{diag}\Big(1 - \big(\mathbf{h}^{(t)}\big)^2\Big)\nabla_{\mathbf{h}^{(t)}} L \tag{63}$$

We now compute the gradient for the weights $\mathbf{V}$. Changing $\mathbf{V}$ only effects that time step and doesn't effect the other nodes in the network. We will be interested in finding $\nabla_{\mathbf{V}} L$. To accomplish this we will be working with the expression found in (23) and computing the partial derivatives for each member of $\mathbf{V}$. Hence at a given time step and a particular value of $\mathbf{V}$ (namely v_{11}) we have

$$L^{(t)} = -\ln\big(\hat{y}_1^{(t)}\big) = -o_1^{(t)} + \ln\Big(\sum_{i=1}^{3} \exp\big(o_i^{(t)}\big)\Big) \tag{64}$$

$$= -c_1 - \sum_{i=1}^{5} v_{1i} h_i^{(t)} + \ln\Big(\sum_{k=1}^{3} \exp\big(c_k + \sum_{l=1}^{5} v_{kl} h_l^{(t)}\big)\Big) \tag{65}$$

then

$$\frac{\partial L^{(t)}}{\partial v_{11}} = -h_1^{(t)} + \frac{h_1^{(t)} \exp\big(c_k + \sum\limits_{l=1}^{5} v_{kl} h_l^{(t)}\big)}{\sum\limits_{k=1}^{3} \exp\big(c_k + \sum\limits_{l=1}^{5} v_{kl} h_l^{(t)}\big)} \tag{66}$$

$$= -h_1^{(t)} + h_1^{(t)} \hat{y}_1^{(t)} \tag{67}$$

$$= h_1^{(t)}\big(\hat{y}_1^{(t)} - 1\big) \tag{68}$$

repeating the same for $v_{1i}, i = 2, \ldots, 5$

$$\frac{\partial L^{(t)}}{\partial v_{12}} = h_2^{(t)}\big(\hat{y}_1^{(t)} - 1\big) \tag{69}$$

$$\frac{\partial L^{(t)}}{\partial v_{13}} = h_3^{(t)}\big(\hat{y}_1^{(t)} - 1\big) \tag{70}$$

$$\frac{\partial L^{(t)}}{\partial v_{14}} = h_4^{(t)}\big(\hat{y}_1^{(t)} - 1\big) \tag{71}$$

$$\frac{\partial L^{(t)}}{\partial v_{15}} = h_5^{(t)}\big(\hat{y}_1^{(t)} - 1\big) \tag{72}$$

Continuing for the remaining values of $\mathbf{V}$ we arrive at the following

$$\begin{pmatrix} \hat{y}_1^{(t)} - 1 \\ \hat{y}_2^{(t)} \\ \hat{y}_3^{(t)} \end{pmatrix} \begin{pmatrix} h_1^{(t)} & h_2^{(t)} & h_3^{(t)} & h_4^{(t)} & h_5^{(t)} \end{pmatrix} \tag{73}$$

$$= \big(\nabla_{\mathbf{o}^{(t)}} L\big)\big(\mathbf{h}^{(t)}\big)^T \tag{74}$$

We turn our attention to computing $\nabla_{\mathbf{W}} L$. Since a perturbation of $\mathbf{W}$ propagates in time we compute $\mathbf{W}$ in the same manner that we did for $\nabla_{\mathbf{h}^{(t)}} L$ which begins at (33). Since the computation is nearly identical to that found in (33) we omit most of it. Beginning at the the final time step τ and writing L as in (37)

$$= -c_1 - \sum_{i=1}^{5} v_{1i} \tanh\Big(b_i + \sum_{\alpha=1}^{5} w_{i\alpha} h_\alpha^{(\tau-1)} + F(\mathbf{x}^{(\tau)})\Big) \tag{75}$$

$$+\ln\left[\sum_{k=1}^{3}\exp\left(c_k+\sum_{l=1}^{5}v_{kl}\tanh\left(b_l+\sum_{\beta=1}^{5}w_{l\beta}h_\beta^{(\tau-1)}+F(\mathbf{x}^{(\tau)})\right)\right)\right]$$

and

$$\begin{aligned}\frac{\partial L^{(\tau)}}{\partial w_{11}} =& -h_1^{(\tau-1)}\big(1-(h_1^{(\tau)})^2\big)\sum_{i=1}^{5}v_{1i} && (76)\\ &+\frac{h_1^{(\tau-1)}\big(1-(h_1^{(\tau)})^2\big)\exp\left(c_1+\sum_{l=1}^{5}v_{1l}\tanh\left(b_l+\sum_{\beta=1}^{5}w_{l\beta}h_\beta^{(\tau-1)}+F(\mathbf{x}^{(\tau)})\right)\right)\sum_{i=1}^{5}v_{1i}}{\sum_{k=1}^{3}\exp\left(c_k+\sum_{l=1}^{5}v_{kl}\tanh\left(b_l+\sum_{\beta=1}^{5}w_{l\beta}h_\beta^{(\tau-1)}+F(\mathbf{x}^{(\tau)})\right)\right)}\\ =& -h_1^{(\tau-1)}\big(1-(h_1^{(\tau)})^2\big)\sum_{i=1}^{5}v_{1i}+\hat{y}_1^{(\tau)}h_1^{(\tau-1)}\big(1-(h_1^{(\tau)})^2\big)\sum_{i=1}^{5}v_{1i} && (77)\\ =& big(1-(h_1^{(\tau)})^2\big)\sum_{i=1}^{5}v_{1i}\big(\hat{y}_1^{(\tau)}-1\big)h_1^{(\tau-1)} && (78)\end{aligned}$$

For the final time step we then arrive at

$$\underbrace{\begin{pmatrix}1-(h_1^{(\tau)})^2 & 0 & 0 & 0 & 0\\ 0 & 1-(h_2^{(\tau)})^2 & 0 & 0 & 0\\ 0 & 0 & 1-(h_3^{(\tau)})^2 & 0 & 0\\ 0 & 0 & 0 & 1-(h_4^{(\tau)})^2 & 0\\ 0 & 0 & 0 & 0 & 1-(h_5^{(\tau)})^2\end{pmatrix}}_{\operatorname{diag}\left(1-\left(\mathbf{h}^{(\tau)}\right)^2\right)} \tag{79}$$

$$\underbrace{\underbrace{\begin{pmatrix}v_{11} & v_{21} & v_{31}\\ v_{12} & v_{22} & v_{32}\\ v_{13} & v_{23} & v_{33}\\ v_{14} & v_{24} & v_{34}\\ v_{15} & v_{25} & v_{35}\end{pmatrix}}_{\mathbf{V}^T}\underbrace{\begin{pmatrix}\hat{y}_1^{(t)}-1\\ \hat{y}_2^{(t)}\\ \hat{y}_3^{(t)}\end{pmatrix}}_{\nabla_{\mathbf{o}}^{(\tau)}L}}_{\nabla_{\mathbf{h}^{(\tau)}}L}\begin{pmatrix}h_1^{(\tau-1)} & h_1^{(\tau-1)} & h_3^{(\tau-1)} & h_4^{(\tau-1)} & h_5^{(\tau-1)}\end{pmatrix}$$

For the second to last time step we would arrive at:

$$\operatorname{diag}\Big(1-(\mathbf{h}^{(\tau-1)})^2\Big)\underbrace{\Big(\mathbf{W}^T\big(\nabla_{\mathbf{h}^{(\tau)}}L\big)\operatorname{diag}\Big(1-(\mathbf{h}^{(\tau)})^2\Big)+\mathbf{V}^T\big(\nabla_{\mathbf{o}^{(\tau)}}L\big)\Big)}_{\nabla_{\mathbf{h}^{(\tau-1)}}L}\big(\mathbf{h}^{(\tau-1-1)}\big) \tag{80}$$

Using the expression in (80) we can back-propagate to $t = 1$. The gradient of the total loss with respect to $\mathbf{W}$ can be expressed as

$$\nabla_{\mathbf{W}}L=\sum_{t}^{\tau}\operatorname{diag}\Big(1-\big(\mathbf{h}^{(t)}\big)^2\Big)\big(\nabla_{\mathbf{h}^{(t)}}L\big)\big(\mathbf{h}^{(t-1)}\big)^T \tag{81}$$

Finally, we compute the change for the input matrix $\mathbf{U}$. This is identical to computing $\nabla_{\mathbf{W}}L$. We notice that the only appearance of $\mathbf{U}$ is in $\mathbf{h}^{(t)}$. This implies that we must simply compute $\nabla_{\mathbf{U}^{(t)}} h_i^{(t)}$ which is done exactly like for $\nabla_{\mathbf{W}} L$ so that

$$\nabla_{\mathbf{U}} L = \sum_{t}^{\tau} \underbrace{\sum_{i} \left(\frac{\partial L}{\partial h_i^{(t)}}\right)}_{\mathrm{diag}\left(1-\left(\mathbf{h}^{(t)}\right)^2\right)} \underbrace{\nabla_{\mathbf{U}} h_i^{(t)}}_{(\mathbf{x}^{(t)})^T} \tag{82}$$

$$= \sum_{t}^{\tau} \mathrm{diag}\left(1 - \left(\mathbf{h}^{(t)}\right)^2\right) \left(\nabla_{\mathbf{h}^{(t)}} L\right) \left(\mathbf{x}^{(t)}\right)^T \tag{83}$$

Summarizing, the gradient for the desired parameters is (with tanh activation and softmax output):

$$\boxed{\begin{aligned}
\nabla_{\mathbf{c}} L &= \sum_{t}^{\tau} \left(\frac{\partial \mathbf{o}^{(t)}}{\partial \mathbf{c}}\right)^T \nabla_{\mathbf{o}^{(t)}} L = \sum_{t}^{\tau} \nabla_{\mathbf{o}^{(t)}} L && (84)\\
\nabla_{\mathbf{b}} &= \sum_{t}^{\tau} \left(\frac{\partial \mathbf{h}^{(t)}}{\partial \mathbf{b}^{(t)}}\right)^T \nabla_{\mathbf{h}^{(t)}} L = \sum_{t}^{\tau} \mathrm{diag}\left(1 - \left(\mathbf{h}^{(t)}\right)^2\right) \nabla_{\mathbf{h}^{(t)}} L && (85)\\
\nabla_{\mathbf{V}} L &= \sum_{t}^{\tau} \sum_{i} \left(\frac{\partial L}{\partial o_i^{(t)}}\right) \nabla_{\mathbf{V}} o_i^{(t)} = \sum_{t}^{\tau} \left(\nabla_{\mathbf{o}^{(t)}} L\right) \left(\mathbf{h}^{(t)}\right)^T && (86)\\
\nabla_{\mathbf{W}} L &= \sum_{t}^{\tau} \sum_{i} \left(\frac{\partial L}{\partial h_i^{(t)}}\right) \nabla_{\mathbf{W}} h_i^{(t)} = \sum_{t}^{\tau} \mathrm{diag}\left(1 - \left(\mathbf{h}^{(t)}\right)^2\right) \left(\nabla_{\mathbf{h}^{(t)}} L\right) \left(\mathbf{h}^{(t-1)}\right)^T && (87)\\
\nabla_{\mathbf{U}} L &= \sum_{t}^{\tau} \sum_{i} \left(\frac{\partial L}{\partial h_i^{(t)}}\right) \nabla_{\mathbf{U}} h_i^{(t)} = \sum_{t}^{\tau} \mathrm{diag}\left(1 - \left(\mathbf{h}^{(t)}\right)^2\right) \left(\nabla_{\mathbf{h}^{(t)}} L\right) \left(\mathbf{x}^{(t)}\right)^T && (88)
\end{aligned}}$$

[1]

References

[1] Ian Goodfellow Yoshua Bengio and Aaron Courville. Deep learning. Book in preparation for MIT Press, 2016.

[2] Alex Graves. Supervised sequence labelling with recurrent neural networks, 2010.

[3] Raul Rojas. Neural Networks - A Systematic Introduction. Springer, 1996.