Exercise 3 - Recurrent Networks and NLP

Abdalla Arafa, 428044
Ehsan Attar, 427214
Sai Leela Poduru, 428272
Prateek Rathod, 428396
Shruti Shrivastava, 428364
Erik Schwede, 428240

December 2023

3.1. Backpropagation through Time

i)

Show that:

$$\frac{\partial L}{\partial W} = \sum_{t=1}^{T} \sum_{k=1}^{t} \frac{\partial L_t}{\partial h_t} \frac{\partial h_t}{\partial h_k} \frac{\partial h_k}{\partial W}$$

The loss functions is defined as:

$$L = \sum_{t=1}^{T} L_t$$

$$\frac{\partial L}{\partial W} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial W}$$

Lets expand the sum:

$$\frac{\partial L_t}{\partial W} = \frac{\partial L_t}{\partial h_t} \cdot \left(\frac{\partial h_t}{\partial W} + \frac{\partial h_t}{\partial h_{t-1}} \cdot \frac{\partial h_{t-1}}{\partial W} + \cdots + \frac{\partial h_t}{\partial h_1} \cdot \frac{\partial h_1}{\partial W}\right)$$

$$\frac{\partial L_t}{\partial W} = \frac{\partial L_t}{\partial h_t} \sum_{k=1}^{t} \frac{\partial h_t}{\partial h_k} \cdot \frac{\partial h_k}{\partial W}$$

$$\frac{\partial L}{\partial W} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial h_t} \sum_{k=1}^{t} \frac{\partial h_t}{\partial h_k} \cdot \frac{\partial h_k}{\partial W}$$

$$\frac{\partial L}{\partial W} = \sum_{t=1}^{T} \sum_{k=1}^{t} \frac{\partial L_t}{\partial h_t} \frac{\partial h_t}{\partial h_k} \frac{\partial h_k}{\partial W}$$

ii)

Given:

$$f(h) = \sigma(W \cdot h)$$

Show that:

$$\frac{\partial f}{\partial h} = diag(\sigma'(Wh))W$$

The partial derivative for the i-th value of f:

$$\frac{\partial f_i}{\partial h} = \sigma'(W_i^T \cdot h) \cdot W_i^T$$

$$\frac{\partial f_i}{\partial h} = \sigma'(W_i^T \cdot h) \cdot [w_{i,0}, w_{i,1}, \cdots, w_{i,n}]$$

$$s_i = \sigma'(W_i^T \cdot h)$$

$$\frac{\partial f_i}{\partial h} = s_i \cdot [w_{i,0}, w_{i,1}, \cdots, w_{i,n}]$$

$$\frac{\partial f_i}{\partial h} = [w_{i,0} \cdot s_i, w_{i,1} \cdot s_i, \cdots, w_{i,n} \cdot s_i]$$

$$\frac{\partial f}{\partial h} = \begin{bmatrix} w_{0,0} \cdot s_0 & w_{0,1} \cdot s_0 & \cdots & w_{0,n} \cdot s_0 \\ w_{1,0} \cdot s_1 & w_{1,1} \cdot s_1 & \cdots & w_{1,n} \cdot s_1 \\ \vdots & \vdots & \vdots & \vdots \\ w_{i,0} \cdot s_i & w_{i,1} \cdot s_i & \cdots & w_{i,n} \cdot s_i \end{bmatrix}$$

$$\frac{\partial f}{\partial h} = \begin{bmatrix} s_0 & 0 & \cdots & \cdots & 0 \\ 0 & s_1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \cdots & 0 & s_i \end{bmatrix} \cdot \begin{bmatrix} w_{0,0} & w_{0,1} & \cdots & w_{0,n} \\ w_{1,0} & w_{1,1} & \cdots & w_{1,n} \\ \vdots & \vdots & \vdots & \vdots \\ w_{i,0} & w_{i,1} & \cdots & w_{i,n} \end{bmatrix}$$

$$\frac{\partial f}{\partial h} = diag(\sigma'(Wh))W$$

iii)

As proven in ii) if $f(h) = \sigma(Wh)$ then $\frac{\partial f}{\partial h} = diag(\sigma'(Wh))W$.
That means that $\frac{\partial h_t}{\partial h_{t-1}}$ is on the shape of $diag(\sigma')W$.
The extended sum for T=3 equals to:

$$\frac{\partial L}{\partial W} = \sum_{t=1}^{3} \sum_{k=1}^{t} \frac{\partial L_t}{\partial h_t} \frac{\partial h_t}{\partial h_k} \frac{\partial h_k}{\partial W}$$

$$\frac{\partial L}{\partial W} = \frac{\partial L_1}{\partial h_1} \frac{\partial h_1}{\partial h_1} \frac{\partial h_k}{\partial W} + \frac{\partial L_2}{\partial h_1} \frac{\partial h_2}{\partial h_2} \frac{\partial h_1}{\partial W} + \frac{\partial L_2}{\partial h_2} \frac{\partial h_2}{\partial h_2} \frac{\partial h_2}{\partial W} + \frac{\partial L_3}{\partial h_3} \frac{\partial h_3}{\partial h_3} \frac{\partial h_3}{\partial W} + \frac{\partial L_3}{\partial h_3} \frac{\partial h_3}{\partial h_2} \frac{\partial h_2}{\partial W} + \frac{\partial L_3}{\partial h_t} \frac{\partial h_3}{\partial h_2} \frac{\partial h_2}{\partial h_1} \frac{\partial h_1}{\partial W}$$

That means that $\frac{\partial h_t}{\partial h_{t-1}} = diag(\sigma')W$
As it can be observed in the extended sum there are two matrices getting multiplied(marked in red). Which corresponds with the the rule the $T-1$ matrixes of type $\frac{\partial h_i}{\partial h_{i-1}} \cdot \frac{\partial h_j}{\partial h_{j-1}}$ should be multiplied.
In case of an arbitrary T:

$$\text{for } i > j$$

$$\frac{\partial h_i}{\partial h_j} = \prod_{k=j}^{i-1} \frac{\partial h_{k+1}}{\partial h_k}$$

$$\frac{\partial h_k}{\partial h_1} = \prod_{k=1}^{T-1} \frac{\partial h_{k+1}}{\partial h_k}$$

The number of elements of this product can be calculated as follows:

$$T - 1 - 1 + 1 = T - 1$$

iv)

$$A^{30} = Q\Lambda^{30}Q^{-1}$$

$$\Lambda^{30} = \begin{bmatrix} 0.4^{30} & 0 \\ 0 & 0.9^{30} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0.04 \end{bmatrix}$$

$$A^{30} = \begin{bmatrix} 0.014 & -0.019 \\ -0.019 & 0.026 \end{bmatrix}$$

When all the gradients are less then 1 there will be a vanishing gradient problem. If one of them is over 1 we have a exploding gradient problem. If all are equal to one there is the perfect scenario. The loss value will be propagated through all layers of the network as it is.

3.2. Gated recurrent units

i)

Given:

$$u_t = \sigma(w.h_{t-1} + w.x_t) \tag{1}$$

$$s_t = w.(h_{t-1} + x_t) \tag{2}$$

$$h_t = u_t.h_{t-1} + (1 - u_t).s_t \tag{3}$$

Calculating derivative of s from equation 2:

$$(4)$$

$$s_t = w.(h_{t-1} + x_t) \tag{5}$$

$$\frac{\partial s_t}{\partial h_{t-1}} = w.\frac{\partial h_{t-1}}{\partial h_{t-1}} + 0 \tag{6}$$

$$= w \tag{7}$$

Let's differentiate equation (3) with respect to h_{t-1}:

$$\frac{\partial h_t}{\partial h_{t-1}} = u_t + \sigma'(w.h_{t-1} + w.x_t)h_{t-1} + (1-u_t)w - s_t.w.\sigma'(w.h_{t-1} + w.x_t) \tag{8}$$

$$= w(\sigma'(w.h_{t-1} + w.x_t)h_{t-1} + (1-u_t) - s_t.\sigma'(w.h_{t-1} + w.x_t) + u_t \tag{9}$$

Since we know about the derivative of the sigmoid function -

$$\sigma' = \sigma(1-\sigma) \tag{10}$$

On substituting it back to our previous equation, we have:

$$\frac{\partial h_t}{\partial h_{t-1}} = w.(u_t(1-u_t)h_{t-1} + (1-u_t) - s_t.u_t(1-u_t) + u_t \tag{11}$$

$$= w(1-u_t)(h_{t-1}.u_t - s_t.u_t + 1) + u_t \tag{12}$$

On comparing it with the required form A_t . w + B_t , we get:

$$A_t = (1-u_t)(h_{t-1}u_t - s_t.u_t + 1) \tag{13}$$

$$B_t = u_t \tag{14}$$

ii)

The long-term derivative can be written as:

$$\frac{\partial h_t}{\partial h_0} = \frac{\partial h_t}{\partial h_{t-1}}.\frac{\partial h_{t-1}}{\partial h_{t-2}}.\frac{\partial h_{t-2}}{\partial h_{t-3}}.....\frac{\partial h_1}{\partial h_0} \tag{15}$$

$$or \tag{16}$$

$$\frac{\partial h_t}{\partial h_0} = \prod_{i=0}^{t} \frac{\partial h_i}{\partial h_{i-1}} \tag{17}$$

From part (i) we have found that, for every term:

$$\frac{\partial h_t}{\partial h_{t-1}} = w(1-u_t)(h_{t-1}.u_t - s_t.u_t + 1) + u_t \tag{18}$$

u_t is the output of a sigmoid function that has a value in the range of (0,1). If the value of u_t is close to zero, the weight term won't allow it to disappear. This helps to avoid the vanishing gradient problem, as no exponential decay term can cause the gradients to vanish over long sequences.
When u_t is 1,the B_t function (u_t) avoids the vanishing gradient, it implies that the new hidden state retains the value of the previous hidden state and when u_t is 0, it ignores the information from the previous step.