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:

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

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

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

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

ii)

Given:

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

Show that:

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

iii)

iv)

3.2. Gated recurrent units

i)

Given:

$$u_t = \sigma(w.h_{t-1} + w.x_t) \tag{1}$$
$$s_t = w.(h_{t-1} + x_t) \tag{2}$$
$$h_t = u_t.h_{t-1} + (1 - u_t).s_t \tag{3}$$

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

$$\frac{\partial h_t}{\partial h_{t-1}} = \frac{u_t * h_{t-1} + (1 - u_t) * s_t}{\partial h_{t-1}} \tag{4}$$

On applying the chain rule :

$$\frac{\partial h_t}{\partial h_{t-1}} = u_t.\frac{\partial h_{t-1}}{\partial h_{t-1}} + (1 - u_t).\frac{\partial s_t}{\partial h_{t-1}} \tag{5}$$

We have the value of s from equation 2:

$$\tag{6}$$
$$s_t = w.(h_{t-1} + x_t) \tag{7}$$
$$\frac{\partial s_t}{\partial h_{t-1}} = w.\frac{\partial h_{t-1}}{\partial h_{t-1}} + 0 \tag{8}$$
$$= w \tag{9}$$

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

$$\frac{\partial h_t}{\partial h_{t-1}} = u_t + (1 - u_t).w \tag{10}$$

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

$$A_t = 1 - u_t \tag{11}$$
$$B_t = u_t \tag{12}$$

ii)

The long-term derivative can be written as:

$$\frac{\partial h_t}{\partial h_0} = \frac{\partial h_t}{\partial h_{t-1}} . \frac{\partial h_{t-1}}{\partial h_{t-2}} . \frac{\partial h_{t-2}}{\partial h_{t-3}} ...... \frac{\partial h_1}{\partial h_0} \tag{13}$$

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

$$\frac{\partial h_t}{\partial h_{t-1}} = (1 - u_t).w + u_t \tag{14}$$

u_t is the output of a sigmoid function that has a value in the range of (0,1). Hence, the term remains non-zero. 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, 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.