Mohamed Ehab Fathy 2002597

Hi Star

Assignment 3

1

1) $h_1 = w_x x_1 + w_h * h_0 = 0.1 * 10 + 1 * 1$
$= 2$

$h_2 = w_h * h_1 + w_x * x_2 = 1 * 2 + 0.1 * 10$
$= 3$

$\hat{y}_2 = w_y * h_2 = 2 * 3 = 6$ #

2) $L_t \Longrightarrow$ For 2 outputs

$\hat{y}_2 = 6$

$\hat{y}_1 = w_h * h_1 = 2$

$L_t = \sum_{i=1}^{2} (\hat{y}_i - y_i) = 1 + 9 = 10$

PAGE
DATE

Scanned with CamScanner

3) $$\frac{\partial L_t}{\partial h_1} = \frac{\partial L_1}{\partial h_1} + \frac{\partial L_2}{\partial h_1}$$

* $$\frac{\partial L_1}{\partial h_1} = \frac{\partial L_1}{\partial y_1} * \frac{\partial y_1}{\partial h_1}$$

$$= 2(\hat{y}_1 - y_1) \times w_y$$

$$= -6 * 2 = -12$$

* $$\frac{\partial L_2}{\partial h_1} = \frac{\partial L_2}{\partial y_2} * \frac{\partial y_2}{\partial h_2} * \frac{\partial h_2}{\partial h_1}$$

$$= 2(\hat{y}_2 - y_2) \times w_y \times w_h$$

$$= 2 * 2 \times 1 = 4$$

$$\therefore \frac{\partial L_t}{\partial h_1} = -8$$

4) $$\frac{\partial L_t}{\partial w_h} = \frac{\partial L_1}{\partial w_h} + \frac{\partial L_2}{\partial w_h}$$

$$* \quad \frac{\partial L_1}{\partial w_h} = \frac{\partial L_1}{\partial y_1} * \frac{\partial y_1}{\partial h_1} * \frac{\partial h_1}{\partial w_h}$$

$$= 2(\hat{y}_1 - y_1) * w_y * h_0$$

$$= -6 * 2 * 1 = -12$$

$$* \quad \frac{\partial L_2}{\partial w_h} = \sum_{i=2}^{0} \frac{\partial L_2}{\partial y_2} * \frac{\partial y_2}{\partial h_2} \frac{\partial h_2}{\partial h_i} * \frac{\partial h_i}{\partial w_h}$$

$$= 2(\hat{y}_2 - y_2) * w_y \left[\frac{\partial h_2}{\partial w_h} + \frac{\partial h_2}{\partial h_1} * \frac{\partial h_1}{w_h} + \frac{\partial h_2}{\partial h_0} * \frac{\partial h_0}{\partial w_h}\right]$$

$$= 2(\hat{y}_2 - y_2) * w_y [h_1 + w_h * h_0]$$

$$= 2 * 2 * [2 + 1] = 8$$

$$\therefore \frac{\partial L_t}{\partial w_n} = -4$$

$\boxed{2}$ * to observe this, we must check the impact of the First input on the last outpt

$$\frac{\partial J_n}{\partial w} = \sum_{i=0}^{n} \frac{\partial J_n}{\partial y_n} * \frac{\partial y_n}{\partial h_n} * \frac{\partial h_n}{\partial h_i} * \frac{\partial h_i}{\partial w}$$

@ $i=0$ } ⟶ First input

$$\frac{\partial h_n}{\partial h_0} = \frac{\partial h_n}{\partial h_{n-1}} * \frac{\partial h_{n-1}}{\partial_{n-2}} * \cdots * \frac{\partial h_1}{\partial h_0}$$

and $\frac{\partial h_n}{\partial_{n-1}} = w_{hh} * \tanh'(w_{hh} h_{n-1}, w_{xh} x_n)$

* where w_{hh} is probably sampled From unit

gaussian (< 1), and deriative of tanh ≤ 1

, so we Keep multiplying small numbers

, so $\frac{\partial h_n}{\partial h_0}$ becomes very small no., hence

h_0 has no impact on y_n (Vanishing gradients) (short memory)

$\boxed{3}$ For long sequences, where Vanishing gradients must be avoided, so we can Keep track of long term info.

$\boxed{4}$ Adv: using as training algorithm which is applied to sequence data, where RNN is used (since weights are shared for each time step, then errors are calculated and accumlated for each time step to update single weight)

* disadv: high computational cost for single paramater update

5) a) feed forward NN will assume that prediction of each letter depends only on that letter, which isn't the case, since we have to learn from the whole sequence before predicting one letter

b) word embedding where each letter is encoded to hot vector

c) many to many (encoder - decoder) since we need to watch the whole sequence before output the first letter

e) * choose similar length texts to be in the same patch

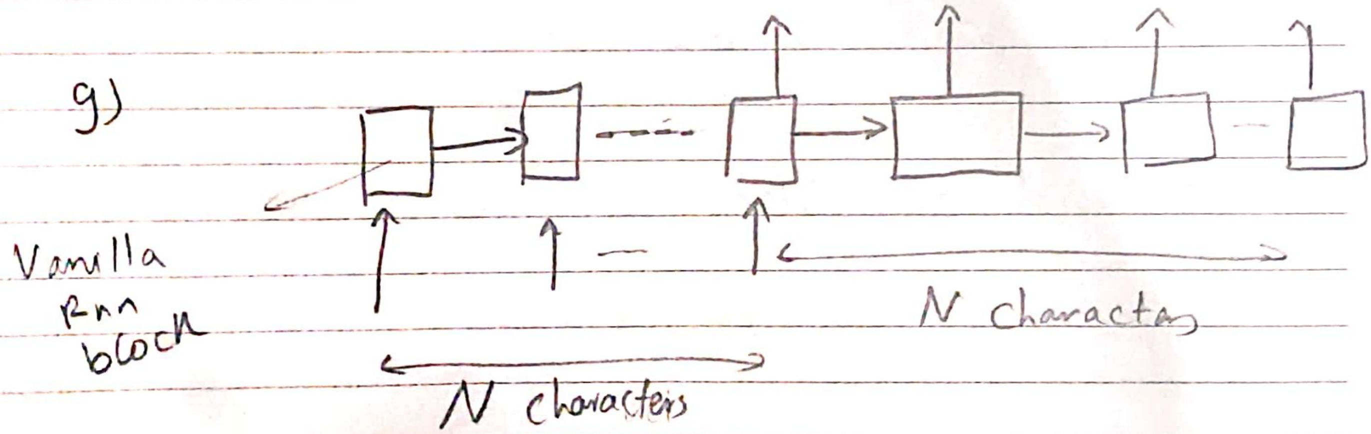
* use padding

* or choose patch size of 1

f) Convert each charchter into hot vector

* divide training data into patches, same text lengths per patch

g)



* encoder decoder arch

h. decode the hot vector into characters
the Concatanate them