CS 577 - F22 - Assignment 5 Questions

Due date: 11/22/2022

Anas Puthawala - A20416308
Professor Gady Agam

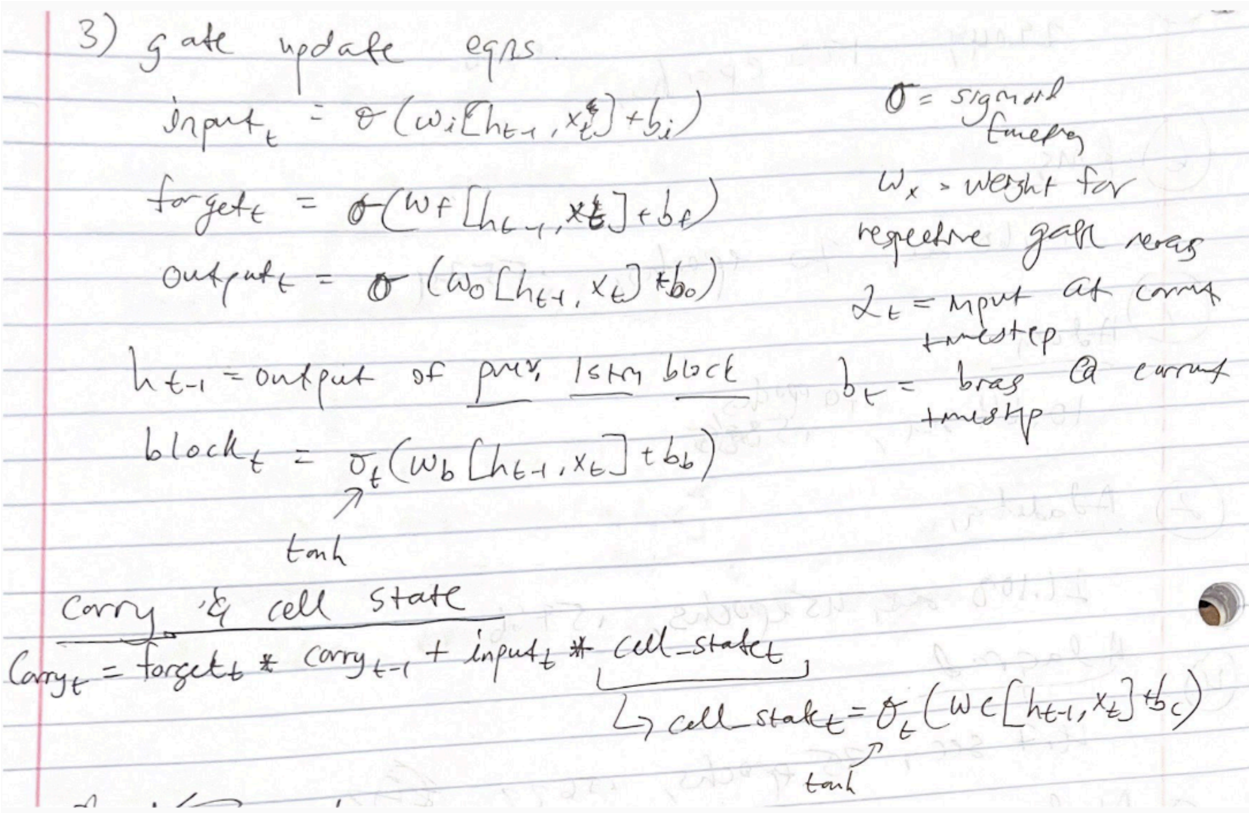
1. The quick brown fox jumped over the lazy dog
9 total words.

The:	[1,0,0,0,0,0,0,0,0]
Quick:	[0,1,0,0,0,0,0,0,0]
Brown:	[0,0,1,0,0,0,0,0,0]
Fox:	[0,0,0,1,0,0,0,0,0]
Jumped:	[0,0,0,0,1,0,0,0,0]
Over:	[0,0,0,0,0,1,0,0,0]
The:	[1,0,0,0,0,0,0,0,0]
Lazy:	[0,0,0,0,0,0,0,1,0]
Dog:	[0,0,0,0,0,0,0,0,1]

2.

Word	The	Quick	Brown	Fox	Jumped	Over	The	Lazy	Dog
Animal	0	0	0	1	0	0	0	0	1
Color	0	0	1	0	0	0	0	0	0
Action	0	1	0	0	1	0	0	1	0

3.



3) gate update eqns.

$\text{input}_t = \sigma(w_i[h_{t-1}, x_t] + b_i)$

$\text{forget}_t = \sigma(w_f[h_{t-1}, x_t] + b_f)$

$\text{output}_t = \sigma(w_o[h_{t-1}, x_t] + b_o)$

σ = sigmoid function

w_x = weight for respective gate neurons

x_t = input at current timestep

b_t = bias @ current timestep

h_{t-1} = output of prev. lstm block

$\text{block}_t = \sigma_t(w_b[h_{t-1}, x_t] + b_b)$

tanh

Carry & cell state

$\text{Carry}_t = \text{forget}_t * \text{carry}_{t-1} + \text{input}_t * \underbrace{\text{cell\_state}_t}$

↳ $\text{cell\_state}_t = \sigma_t(w_c[h_{t-1}, x_t] + b_c)$

tanh

4. Some of the activation functions are from 0,1 (sigmoid for example) that controls how much information will flow through the respective gates that use sigmoid activation on a scale from 0 to 1. However, the others use tanh and this activation function ranges from -1 to 1. The idea behind this is that we don't want only the function to control the flow of information we ALSO want it to control somewhat the amount of information or if it should be negated or not (i.e. it could be in the negative and we can consider that almost like 'forgetting' some flow of information)

5. Longer term dependencies become a problem due to vanishing gradients / exploding gradients and this problem is resolved by the addition of the various gates each serving its own purpose along with the carry state and cell state which takes into account multiple gates at once. The carry state along with the multiple gates aim to resolve the problem of vanishing gradients and also improve on analyzing properties of long-term dependencies.

6. The need for a bi-directional RNN is that sometimes only analyzing inputs sequentially will not provide the best understanding / representation of that text as opposed to if model could also perform the same processes from the other direction.

An example of when bi-directional RNN can perform better than a simple RNN is using text data because this way we'll be able to perform dependency parsing better (i.e. knowing which term refers to what word in a latter or former part of the sentence).

An example when bi-directional RNN won't perform better than a simple RNN is when using sequence data such as stock prices because often times we want to consider the past and simply be able to look ahead so the sequential nature of a simple RNN is preferred rather than going back and fourth from the bi-directional rNN.

7. In sequence-to-sequence learning has encoder and decoders, the encoder serves to pass condensed information from the input sequence to the decoder, and the decoder aims to utilize that condensed representation of the sequence to output another sequence as intended per task (i.e. translation, captioning, whatever it may be)

In the case of machine translation, the encoder will process the input sequence and have an internal representation of the text, and the decoder will get this internal representation and output words in the target language (i.e. French if we're going from Eng -> French).

The training set for the encoder is simply the english sequence, and the training set for the decoder will be the target sequence (i.e. in this example, French). Also, the target sequence will be shifted forward in the decoder output. The loss we can use can be cross entropy loss.

The role of the encoder is to have an internal condensed representation, the input to the encoder is the sequence in english (again this is our example) and the output may be some internal state / representation of that english text which contains knowledge of the meaning, words used, frequency of words used, dependency of words on other words (example: attention mechanism from transformer network), and more.