CS4851/6851 IDL Homework 6 **April 18, 2022**

Note: All coding problems to be submited with Github Link. Do not Upload the files/folder. Use git commands only.

Note: this is the distribution of questions:

(a) Question 1 to Question 2: Required for everyone.

(b) Question 3 and Question 4: Required by Graduate Students and Bonus for Undergrads

(c) Question 5 to Question 6: Bonus question for both Graduate Students and Undergraduate Students

Problem 1 (10 points)

We can represent the words in a vocabulary with binary vectors that havedimen- sion of the number of words in the vocabulary and all values set to zero except the one value that corresponds to the index of the given word in the sorted version of this vocabulary. This is the so called one-in-K or one-hot encoding.

(a) Describe a representation of a document with a vector. (Think of a represen- tation that is based on the one-hot encoding of the words in that document and has the same dimension as a single word (size of the vocabulary).)

Given a fixed set of vocabulary V = {w1,w2,w3...wn}, one way to represent word w is to encode it with a |V|-dimensional vector, where each dimension is either 0 or 1. This maps each word to an index in the vocabulary. A document equals the summation of words included in a vocabulary dependent of the length of the vocabulary vector.

(b) Explain why this representation is problematic:

(i) Simple sentence or two

- The curse of dimensionality becomes an issue, with just a few words the number of dimensions, the amount of memory is large and costly. Adding in an entire sentence or two would takes a large amount of computation power.

(ii) Examples of the problem(s)

- Example if we have a vocabulary of 25,000, each word is represented by 24,999 zeros and a single one leading to us needing 25,0000 squared = 6.2 million units of memory space.

(c) Provide atleast two more options to fix this problem.

(i) When it comes to the curse of dimensionality, the matrix takes multiple zeros which are essentially just taking up a lot of space unnecessarily. A way to solve this is utilize a structure that separates all the zeros that are not necessarily needed from the algorithm and delete the memory space they take up when it comes to run time.

Problem 2 (30 points)

A recurrent network in Figure 3 takes a sequence of integers as an input and at the end of the sequence, on the last element produces a number between 0 and

1. What does a 0 mean? What does a 1 mean? Describe which function this network is computing (what is the meaning of this function). Assume all biases are 0, and make sure the hidden state is initialized to 0 as well. Note, the inputs, the weights, and the hidden state are just scalars in this RNN.
   This is a many to one type RNN, taking in a sequence of inputs and resulting in one output being submitted. Here $W_{h\_h} = -1$, $W_{x\_h} = 1$ and $W_{h\_y} = 10$.
   $h_t = f_w (h_{t-1}, xt)$, w/h_t = new state, f_w = some function with parameters W, h_{t-1} = old state, x_t = input vector at some time step.
   $y_t = W_{hy}h_t + b_y$, $h_t = f_w (h_{t=1}, xt)$ -> $h_t = \tanh (W_{hh}h_{t-1} + W_{xh}x_t + b_h)$.

   Plugging in $y_t = W_{hy}h_t + b_y$ -> $y_t = W_{hy}[\tanh (W_{hh}h_{t-1}+W_{xh}x_t+b_h)] + b_y$. This results in: Initial Step y_t = (10) [tanh ((-1) (0) + (1) xt + 0)] -> (10) tanh (-1) + x_t -> tanh (-1) = -.76, this brings the result of the output down.

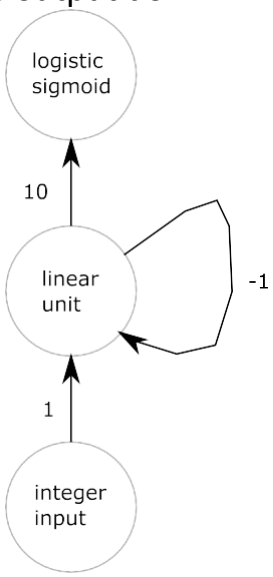

Figure 1: The RNN for Problem 2

Bonus for undergraduates beyond this line.

Problem 3 (20 points)

In this problem, you will implement a recurrent neural network which implements binary addition. The inputs are given as binary sequences, starting with the least significant binary digit. (It is easier to start from the least significant bit, just like how you did addition in grade school.) The sequences will be padded with at least one zero on the end. For instance, the problem

$$100111 + 110010 = 1011001 \quad (1)$$

would be represented as : (a)

Input 1: 1, 1, 1, 0, 0, 1, 0

(b) Input 2: 0, 1, 0, 0, 1, 1, 0

(c) Correct output: 1, 0, 0, 1, 1, 0, 1

There are two input units corresponding to the two inputs, and one output unit. Therefore, the pattern of inputs and outputs for this example would be:

Design the weights and biases for an RNN which has two input units, three hidden units, and one output unit, which implements binary addition. All of the units use the hard threshold activation function. In particular, specify weight

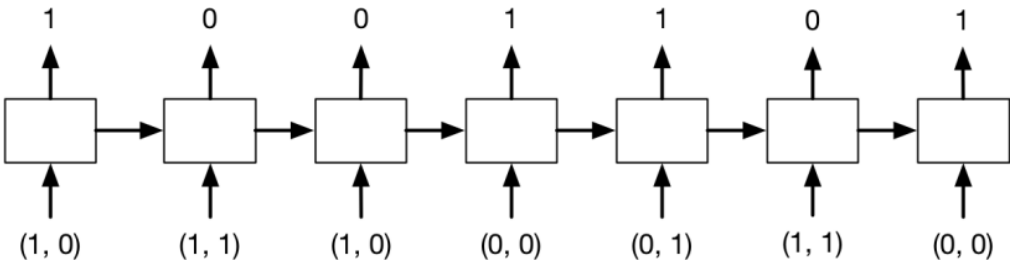

Figure 2: The RNN Binary for Problem 3

matrices U, V, and W, bias vector bh, and scalar bias by for the following archi- tecture:

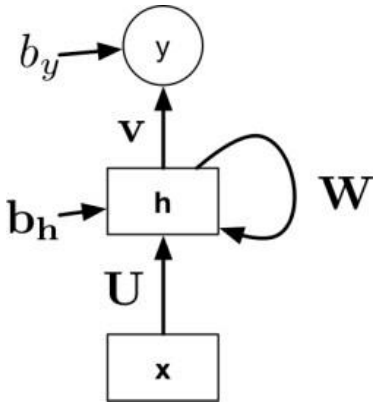

Figure 3: The RNN Architecture for Problem 3

Hint: In the grade school algorithm, you add up the values in each column, including the carry. Have one of your hidden units activate if the sum is at least 1, the second one if it is at least 2, and the third one if it is 3.

$h_t = f_w (h_{t-1}, xt)$, w/h_t = new state, f_w = some function with parameters W, h_{t-1} = old state, x_t = input vector at some time step.

$y_t = W_{hy}h_t + b_y$, $h_t = f_w (h_{t=1}, xt)$ -> $h_t = \tanh (W_{hh}h_{t-1} + W_{xh}x_t + b_h)$.

$a_t = b + Wh_{t-1} + Ux_t$, $h_t = \tanh(a_t)$ $o_t = c + Vh_{t-1}$ o_t = output

Problem 4 (20 points)

We have learned about regularization in image processing. How does regulariza- tion help in the context of Recurrent Neural Networks?

Regularization lowers the complexity of a neural network when it comes to training, thus preventing overfitting. When it comes to RNNs we can reduce the number of hidden layers, the number of layers, applying dropouts or regularizers.

Bonus for both undergraduates and graduates beyond this line.

Problem 5 (40 points)

How is teacher forcing ratio more accurate than the model output for a sequence of inputs? How can we use teacher process to parallelize the computation?

Teacher forcing makes the training process faster by parallelization of learning in different time steps. Ground truth output in the current time step is used to compute the system state in the next time steps.

Problem 6 (40 points)

Write a report on one of the following topics:

(a) Attention Is All You Need {https://arxiv.org/pdf/1706.03762.pdf}

(b) Transformers: {https://arxiv.org/pdf/1910.03771v5.pdf}