Purdue University Duc Viet Le
CS 59000SA

Assignment 5

Problem 1. Assume we have a sequence of t inputs, namely $x_1, x_2, \ldots, x_t$. Construct neural networks based on the following requirements.

1. Draw a feed-forward neural network using A that will perform equivalent computation on the given input as the above RNN. Clearly mark input, output and all connections. Justify your answer.

   Answer. This is an equivalent feed-forward neural network using A that will perform equivalent computation.

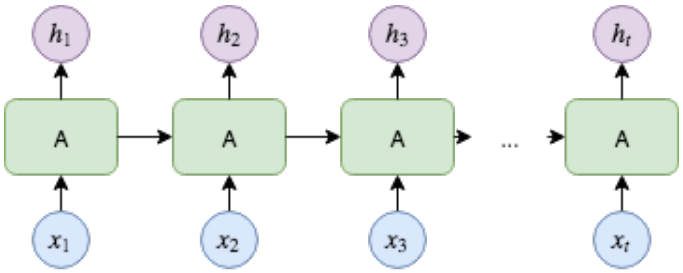

Figure 1: Equivalent feed-forward neural network using A

   x_i just as input at different time.

2. To use a feed-forward neural network to simulate a RNN, what weight constraints are needed? What's the space complexity for the model of RNN? What about it's equivalent feed-forward network version?

   Ans: The weight has to be the same at every layer. Also, we want $w_i = w_j$ for all $i \neq j$, and when we updates the weight values, we want the weight changes has to be the same (i.e. $\Delta w_i = \Delta w_j$).

   If x_i is the scalar value, the space complexity $\mathcal{O}(1)$ for RNN because we only need to store one input at a time. In the equivalent feed-forward network, it will be $\mathcal{O}(t)$

   Otherwise, if x_i is vector (says, size n), then it will be $\mathcal{O}(n)$ for RNN, and $\mathcal{O}(nt)$ for feed-forward.

Problem 2. In this question, you are going to learn how gradient vanish problem can occur.

1. The output range of the sigmoid function is:

   Ans: $(-1, 1)^{|h|}$ (vector of length $|h|$, each value in the interval $(-1, 1)$). I assume the $tanh(\cdot)$ is the non-linearity functioned used.

2. Prove why Equation 5.4 can suffer from vanishing gradient or gradient explosion problem

Ans: As explained in the lecture notes, we have:

$$||\frac{\partial h_i}{\partial h_{i-1}}|| \leq ||W^T||||diag(sigmoid(h_{i+1}))|| \leq \alpha_a \alpha_b \tag{1}$$

Therefore, in the equation (5.4), the quantity:

$$\left|\left| \prod_{k+1}^{t} \frac{\partial h_i}{\partial h_{i-1}} \right|\right| \leq (\alpha_a \alpha_b)^{t-k} \tag{2}$$

this quantity can become very large ($\alpha_a \alpha_b > 1$) or very small ($\alpha_a \alpha_b < 1$). Thus, when the gradient value becomes very large, it causes gradient explosion, and when the gradient becomes very small, it causes vanishing gradient.

Problem 3. Sketch the structure of LSTM based on the above description and following equation, clearly mark what's the input and output for each gate.

Answer: The following figure is the sketched structure of LSTM

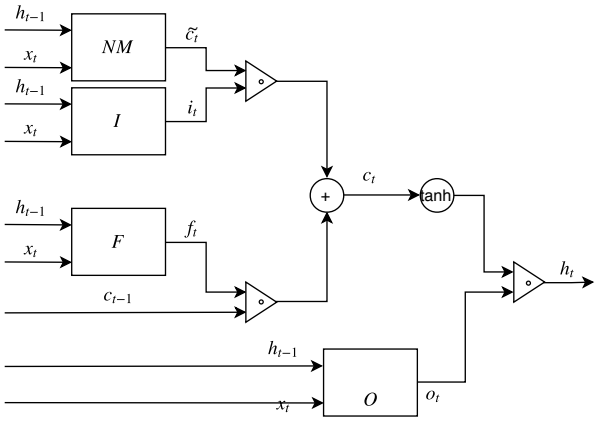

Figure 2: Structure of LSTM

Problem 4. Which gate(s) of LSTM help prevent gradient vanishing problem? Justify your answer.

Ans: Forget gate, $f_t = F(x_t, h_{t-1})$. The idea is that it makes an assessment on wether the past memory cell is useful for the computation of the current memory cell.

Problem 5. What's the difference between a memory cell and a memory block?

Ans: LTSM allows the network to memorize information after several timestep, and those memories are stored in memory cell.

Memory block contains one or more memory cells sharing the same input gate. Memory block of size 1 is the same as a memory cell.