50.021 Artificial Intelligence
Theory Homework 5

Due: every Monday, 4PM before class starts

Q1. Are of the following neural networks suffer from the vanishing gradient problem? Given your reason in not more than 2 sentences.

1. 1-Layer Feed-Foward NN
   Solution:
   There are two factors that affect the magnitude of gradients, which is the weights and the derivatives of the activation functions that the gradient passes through. In this case, no, since its just 1 layer.

2. Very Deep Feed-Forward NN
   Solution:
   Yes, since it is involving chained matrix multiplication and possibly saturating non-linearities..

3. Recurrent NN
   Solution:
   Yes, for the same reason as the previous part, very deep feed-foward NN is just a temporal unfold of recurrent NN.

4. LSTM NN
   Solution:
   No. Since the activation function is identity function, the gradient is approximately 1.

5. GoogleNet
   Solution:
   No. Since the outputs are at 3 different layers.

6. ResNet
   Solution:
   No. ResNet provides shortcut at each layer for sending the gradient signal backwards.

Q2. The following are the defining equations for an LSTM cell,

$$
\begin{aligned}
i_t &= \sigma(W^i x_t + U^i h_{t-1}) \\
f_t &= \sigma(W^f x_t + U^f h_{t-1}) \\
o_t &= \sigma(W^o x_t + U^o h_{t-1}) \\
\tilde{c}_t &= \tanh(W^c x_t + U^c h_{t-1}) \\
c_t &= f_t \circ c_{t-1} + i_t \circ \tilde{c}_t \\
h_t &= o_t \circ \tanh(c_t)
\end{aligned}
$$

The symbol $\circ$ denotes element-wise multiplication and $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function. Answer True/False to the following questions and give not more than 2 sentences explanation.

1. If $x_t = 0$ vector then $h_t = h_{t-1}$.
   Solution:
   False. Sub in the zero vector and after transformation by weight U and σ nonlinearities, $h_t \neq h_{t-1}$.

2. If f_t is very small or zero, then the error will not be back-propagated to earlier time steps.
   Solution:
   False. Because i_t and $\tilde{c}_t$ depends on the previous time steps h_{t-1}.

3. The entries of f_t, i_t, o_t are non-negative.
   Solution:
   True, the σ function ranges between 0 to 1.

4. f_t, i_t, o_t can be seen as probability distributions, which means that their entries are non-negative and their entries sum to 1.
   Solution:
   False, since sigmoid function is applied element-wise, then the entries of those vectors will not necessarily sum up to 1.

Q3. Consider the RNN structure in figure 1. It has a scalar input at the first time step, and make a scalar prediction at the last time step. The activation function at each node is a shifted sigmoid function,

$$g(z) = \sigma(z) - 0.5$$

In other words, the output of each node is multiplied by weights w, and put into the activation function of the next node. For example, $z^2 = g(z^1 \cdot w)$. Answer the following questions,

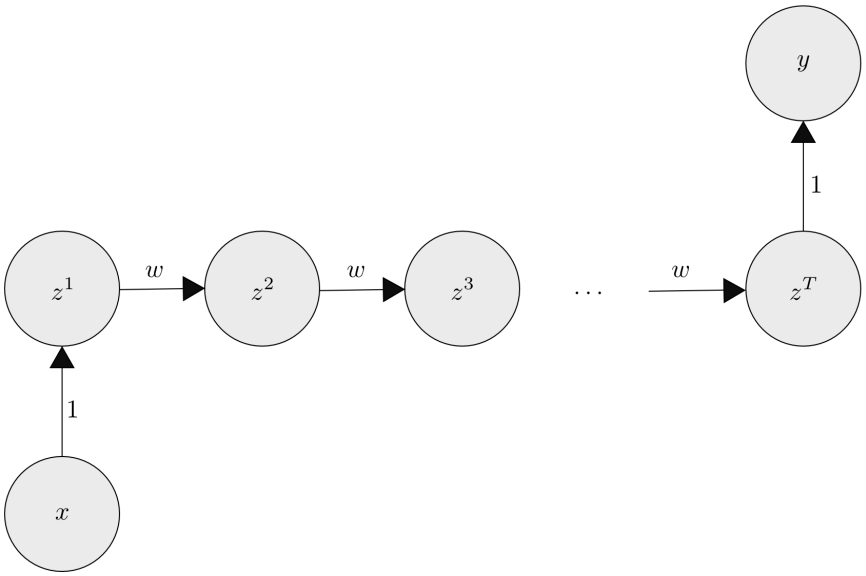


Figure 1: RNN for Q3

1. Write the formula for the error derivative of z^t, $\frac{\partial E}{\partial z^t}$ as a function of the error in the next time step, $\frac{\partial E}{\partial z^{t+1}}$ for $t < T$.
   Solution:

$$z^{t+1} = g(z^t \cdot w) \tag{1}$$

$$\frac{\partial z^{t+1}}{\partial z^t} = \sigma(z^t \cdot w)(1 - \sigma(z^t \cdot w)) \cdot w \tag{2}$$

$$\frac{\partial E}{\partial z^t} = \frac{\partial E}{\partial z^{t+1}} \frac{\partial z^{t+1}}{\partial z^t} \tag{3}$$

$$= \frac{\partial E}{\partial z^{t+1}} \sigma(z^t \cdot w)(1 - \sigma(z^t \cdot w)) \cdot w \tag{4}$$

2. Suppose that the input to the network is $x = 0$. What is the value of z^t for all $t < T$?
   Solution: If $x = 0$ then it is easy to verify that $z^t = 0 \ \forall t$. This is because $\sigma(0) = \frac{1}{2}$, and this value is always subtracted by 0.5 in $g(z)$, resulting in 0 in each neuron.

3. Suppose that the input to the network is $x = 0$ and using your answer in part 2, determine the value β such that if $w < \beta$, the gradient $\frac{\partial E}{\partial z^t}$ vanishes towards zero, and if $w > \beta$, it

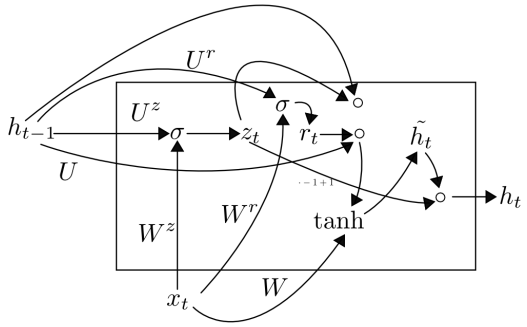
explodes towards infinity, assuming T is large enough.
Solution: One needs to realise that $\sigma(z^t \cdot w)(1 - \sigma(z^t \cdot w)) = \frac{1}{4}$ since $z^t = 0$ for all t. This means that if $w = 4$, then $\frac{\partial E}{\partial z^t} = \frac{\partial E}{\partial z^{t+1}}$. Therefore, $\beta = 4$.

Q4. The defining equations for a GRU cell are,

$$
\begin{aligned}
z_t &= \sigma(W^z x_t + U^z h_{t-1}) \\
r_t &= \sigma(W^r x_t + U^r h_{t-1}) \\
\tilde{h}_t &= \tanh(W x_t + r_t \circ U h_{t-1}) \\
h_t &= z_t \circ h_{t-1} + (1 - z_t) \circ \tilde{h}_t
\end{aligned}
$$

1. Draw the diagram of this GRU cell.

   Solution:



2. Assume h_t and x_t are column vectors, with dimensions d_h and d_x respectively. What are the dimensions (rows $\times$ columns) of the weight matrices W^z, W^r, W, U^z, U^r, and U?

   Solution:

$$
\begin{aligned}
W^z &: d_h \times d_x \\
U^z &: d_h \times d_h \\
W^r &: d_h \times d_x \\
U^r &: d_h \times d_h \\
W &: d_h \times d_x \\
U &: d_h \times d_h
\end{aligned}
$$

3. Like LSTM cells, GRU cells can tackle vanishing or exploding gradient problem too. By taking a look at the formula for LSTM in Q2, what is the main advantage of using GRU cells? Give a max of 5 sentences answer.
   Hint: We expect a qualitative answer (deep math proofs are not required) that comes with an explanation of the answer

   Solution: GRU cells have less weight to train. It is a simplified version of LSTM cells.