**Quiz 7** **Name:** ____________________

**Week 8, Oct/28/2020** **On your left:** ____________________

**CS 280: Fall 2020** **On your right:** ____________________

**Instructor: Xuming He, Lan Xu**

**Instructions:**
Please answer the questions below. Show all your work. This is an open-book test. NO discussion/collaboration is allowed.

**Problem 1.** (10 points) *LSTM*.

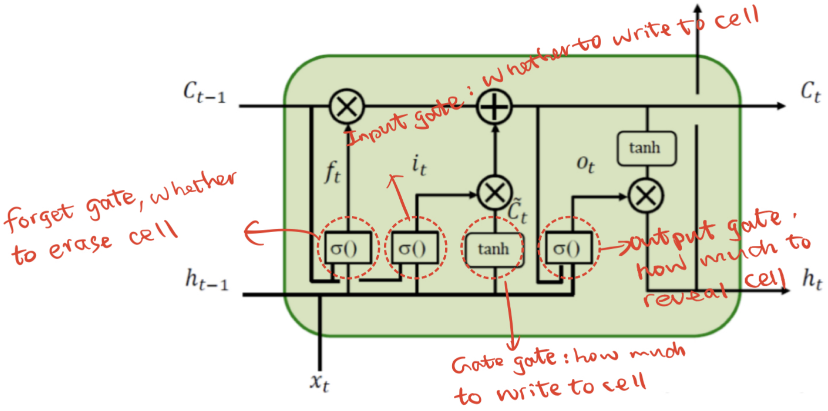
1) Draw the diagram of LSTM, and describe the four gates (where are the gates? what is the purpose/role of each gate?).

2) Why are different activation functions used in different parts of LSTM?

3) Recall that RNNs suffer from vanishing gradients, describe how LSTMs mitigate such problem. (Hint: derive $\frac{\partial C_t}{\partial C_{t-1}}$ similar to Lecture13-Page42 and decide which gate allows the network to better control the gradients)

**Solution:**

1. (5')



$C_{t-1}$
$C_t$
$h_{t-1}$
$h_t$
$x_t$
$f_t$
$i_t$
$o_t$
$\tilde{C}_t$
σ()
σ()
tanh
σ()
tanh
Input gate : whether to write to cell
forget gate, whether to erase cell
output gate, how much to reveal cell
Gate gate : how much to write to cell

2. (2')
sigmoid: squish values between 0 and 1, to forget or keep the value depending on whether it's important.
tanh: ensure that the values stay between -1 and 1, thus regulating the output of the network.

3. (3')

$$\frac{\partial C_t}{\partial C_{t-1}} = \frac{\partial}{\partial C_{t-1}}(f_t * C_{t-1} + i_t * \tilde{C}_t)$$

$$= f_t + \frac{\partial f_t}{\partial C_{t-1}} * C_{t-1} + \frac{\partial i_t}{\partial C_{t-1}} * \tilde{C}_t + i_t * \frac{\partial \tilde{C}_t}{\partial C_{t-1}}$$

It is the presence of the activation value $f_t \in (0, 1)$ from the forget gate along with the additive structure of the gradient equation which allows the LSTMs to better control the gradients to keep important information than RNNs.

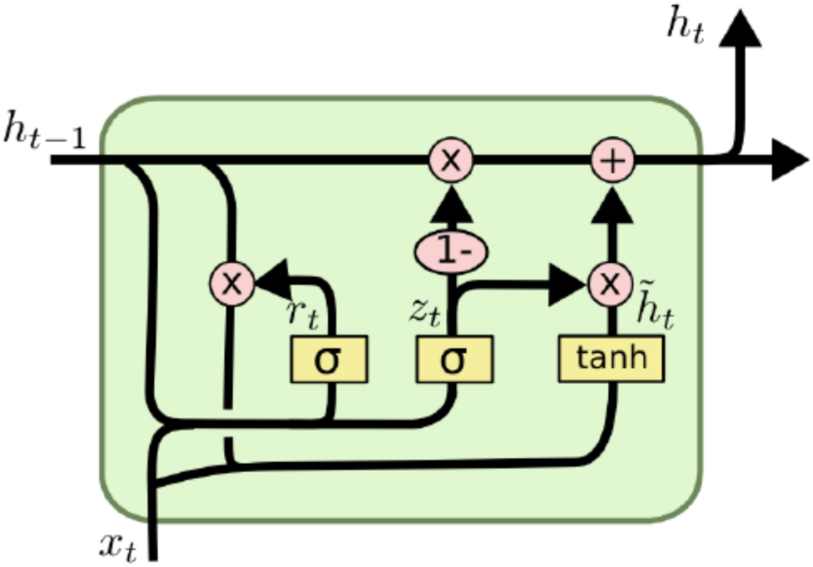
**Problem 2.** (10 points) *GRU*.

1) Draw the diagram of GRU, describe the gates (where? what is the role of each gate?), and point out the differences between GRU and LSTM in the design of gates.

2) In what situation(s) is LSTM/GRU used, respectively? Explain why.

**Solution:**

1. (2') Diagram



$$
\begin{aligned}
z_t &= \sigma\left(W_z \cdot [h_{t-1}, x_t]\right) \\
r_t &= \sigma\left(W_r \cdot [h_{t-1}, x_t]\right) \\
\tilde{h}_t &= \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right) \\
h_t &= (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t
\end{aligned}
$$

2. (3') Reset gate: Combine the forget and input gates in LSTM. Compress memories before using it to decide on the usefulness of the current input.

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

3. (2') Update gate: In this stage, we have two steps: forgetting and remembering. We used the previously obtained update gating to control this process.

$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

4. (3') If there is enough budget and data, LSTM may be better than GRU. But LSTM need more training data and training cost.

   Compared with LSTM, GRU is easier to be trained and can greatly improve the training efficiency. And GRU may perform better when the amount of data is relatively small.