Problem 1: Forward and Backward Propagation (30')

Let's consider a simple two layer neural network.

It has input size 2, one hidden layer size 3, and output size 1.

The input $x = \begin{bmatrix} -2 \\ 3 \end{bmatrix}$, $y = 1$,

two sets of weights $W_1 = \begin{bmatrix} -1.6 & 0.8 \\ 0.3 & 0.6 \\ 1.6 & -0.2 \end{bmatrix}$, $W_2 = \begin{bmatrix} 0.2 & 0.8 & -0.5 \end{bmatrix}$,

biases $b_1 = \begin{bmatrix} 0.3 \\ 0.5 \\ -0.7 \end{bmatrix}$, $b_2 = 0.6$,

all the non-linear activation function is sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$, and loss function $\mathcal{L} = -y \log \hat{y} - (1-y) \log(1-\hat{y})$. (hint: $\sigma'(x) = \sigma(x) * (1 - \sigma(x))$)

1. (2') Calculate $z_1 = W_1 x + b_1$.

2. (4') Calculate $h_1 = \sigma(z_1)$.

3. (4') Calculate $z_2 = W_2 h_1 + b_2$.

4. (4') Calculate $\hat{y} = \sigma(z_2)$.

5. (4') Calculate $\partial \mathcal{L} \ / \ \partial \hat{y}$?

6. (4') Calculate $\partial \hat{y} \ / \ \partial z_2$?

7. (4') Calculate $\partial z_2 \ / \ \partial W_2$?

8. (4') Calculate $\partial \mathcal{L} \ / \ \partial W_2$?

9. (Bonus, 4') Calculate $\partial z_2 \ / \ \partial z_1$?

10. (Bonus, 4') Calculate $\partial z_1 \ / \ \partial W_1$?

11. (Bonus, 4') Calculate $\partial \mathcal{L} \ / \ \partial W_1$?

Problem 2: LSTM (20')

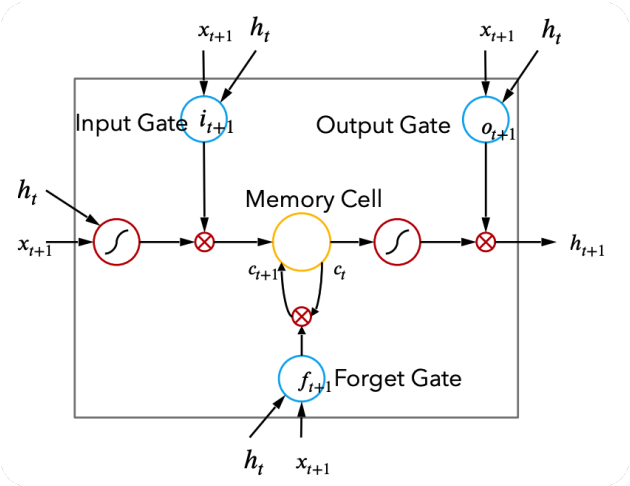

Figure 1: LSTM cell

1. **[10']** Given a LSTM cell in Figure 1, describe the update mechanism of hidden state h_t and cell state c_t by equations. (The weight matrix for the input gate can be indicated by W_{ix} and W_{ih}, and the bias is b_i. Similar for other gates.)

2. **[5']** What is the main difference between LSTM and the vanilla RNN? What advantages does LSTM have?

3. **[5']** Suppose for the t-th step $x_t \in \mathbb{R}^{b \times d_1}, h_t, c_t \in \mathbb{R}^{b \times d_2}$, and there are T timestep for the whole input sequence. Then how many parameters does one LSTM layer have? Here b is the batch size, d_1, d_2 is the dimension. Do not need to consider the word embedding layer.

Problem 3: Transformer (30')

1. **[5']** Write the equation of the scaled dot-product attention in the self-attention layer of Transformer. Here, $Q, K, V \in \mathbb{R}^{b \times d}$ is the query, key, and value. b, d is the batch size and the hidden size.

2. **[5']** What is the purpose of multiple heads in the self-attention layer? What is the purpose of layer normalization?

3. **[5']** Given a sequence $\{x_1, x_2, \cdots, x_n\}$, describe the training objective of GPT model. The model parameter is denoted by θ.

4. **[5']** Suppose we want to get the sentiment for the movie review. There are three sentiment categories: Positive, Negative and Neutral. Is it possible to use GPT3 to do this task? If yes, give a potential task description (for example, the input and output of GPT3). If no, describe the reason.

5. **[10']** What of the following is the most relevant to ChatGPT? Briefly describe the main characteristic of each option and explain your choice.

   (a) ELMo

   (b) BERT

   (c) Roberta

   (d) GPT3

   (e) instructGPT

Problem 4: Regularization (10')

Assume you are training a classification model with 5 output units and the loss function J as defined below. The weight parameters, regularization parameter, and expected and predicted outputs for 5 examples are given below

$$J = \frac{1}{2N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

$$\theta = \begin{bmatrix} 0.2 \\ -0.3 \\ 0.1 \\ 0.5 \\ -0.4 \end{bmatrix}, \lambda = 0.2,$$

$$y_1 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \hat{y_1} = \begin{bmatrix} 0.30 \\ 0.40 \\ 0.10 \\ 0.50 \\ 0.10 \end{bmatrix}, y_2 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \hat{y_2} = \begin{bmatrix} 0.20 \\ 0.30 \\ 0.20 \\ 0.20 \\ 0.10 \end{bmatrix},$$

$$y_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \hat{y_3} = \begin{bmatrix} 0.10 \\ 0.10 \\ 0.60 \\ 0.10 \\ 0.10 \end{bmatrix}, y_4 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \hat{y_4} = \begin{bmatrix} 0.10 \\ 0.50 \\ 0.10 \\ 0.70 \\ 0.10 \end{bmatrix}$$

$$y_5 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \hat{y_5} = \begin{bmatrix} 0.75 \\ 0.10 \\ 0.68 \\ 0.10 \\ 0.60 \end{bmatrix}$$

Redefine your loss function J with

1. **[10']** L2 Regularization and calculate the loss.

Problem 5: Convolutional Layers (10')

1. (10') Suppose we have one batch 100 input images each of size $3 \times 64 \times 64$. Consider a convolutional layer with 2 filters, kernel size 4×4, no padding, and stride of 4. Answer the following questions and given brief explanations for your answers.

   (a) (3') What is the shape of the weight parameters for the convolutional layer?

   (b) (4') What is the output size after we feed the whole batch of input images through the convolutional layer?

   (c) (3') We decide to add a linear layer after the convolutional layer to make a prediction of whether the image is a cat or not, what would be the input dimension for the linear layer?