Instructor: Lei Li

UCSB CS 190I Deep Learning
February 13, 2023

Homework 3

For the following problems, please use LaTex to type your solution. Please submit your solution in PDF. Handwritten solution will not be accepted. You may use the template to write down your solution: https://www.cs.ucsb.edu/~leili/course/dl23w/hw_template.tex.

Problem 1: Multiple Choice (10')

Choose the correct answers from the given candidates

1. (10') What are the meanings of the hyper-parameters η, β_1 and β_2 in Adam respectively?
   - a. Initialized learning rate
   - b. Decay factor of the first derivative
   - c. Decay factor of momentum
   - d. Decay factor of second derivative
   - e. Decay factor of RMSprop
   - f. Final learning rate
   - g. Minimum learning rate
   - h. Maximum learning rate
2. (10') For a binary classification problem with 5 examples, we have:

$$\hat{\mathbf{y}} = \begin{bmatrix} 0.0573 & 0.9427 \\ 0.9852 & 0.0148 \\ 0.4901 & 0.5099 \\ 0.8830 & 0.1170 \\ 0.3686 & 0.6314 \end{bmatrix}, \mathbf{y} = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix}$$

   Here, the $\hat{\mathbf{y}}$ is the output of our model and $\mathbf{y}$ is the label. If we use the cross entropy as the loss function, then the loss should be (one answer):
   - a. 0.2385
   - b. 0.2663
   - c. 0.3198
   - d. 0.4604

Problem 2: Convolutional Neural Networks (40')

1. (10') Given an image of size 64×64, what is the output dimension after applying a convolution with kernel size 5×5, stride of 2, and no padding?

2. (10') For the output above, what is the output dimension after we apply a max pooling layer with the kernel size 2 x 2, stride of 2, and no padding?

3. (10') Given an input of dimension C × H × W , we apply one convolutional layer with kernel size K × K, padding P, stride of S, and F channels. What is the number of parameters for the convolution layer (considering the bias)?

4. (10') How many learnable parameters does one batch normalization layer have?

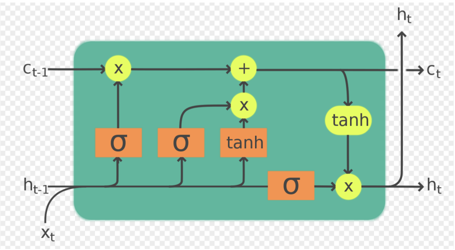

Figure 1: LSTM cell at time step t

Problem 3: Long-Short Term Memory Neural Network (45')

Above picture shows the forward cell of LSTM at time step t, based on which the forward process can be listed as follows:

$$
\begin{aligned}
f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f) \\
i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i) \\
o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o) \\
\tilde{c_t} &= \tanh(W_c x_t + U_c h_{t-1} + b_c) \\
c_t &= f_t * c_{t-1} + i_t * \tilde{c_t} \\
h_t &= o_t * \tanh(c_t)
\end{aligned}
\tag{1}
$$

Suppose $x = \{x_1, x_2, ..., x_L\}$ is a text sequence with L words. x_t is the embedding of t-th word, and the embedding size is 100. Suppose the dimension of h_t is 256.

1. (10') Calculate the size of all the parameters, including W_*, U_* and b_*.
2. (5') If we input a batch of sequences into LSTM with different sequence length, how can we deal with the sequences with variant lengths?
3. (5') Does the final hidden representation h_L include information of the whole sequence x?
4. (5') What are the functions for input gate, forget gate and output gate? Does a larger input gate lead to larger forget gate?
5. (5') Original RNN may have gradient vanishing problem. Can LSTM relieve this issue? Why?
6. (5') Original RNN has gradient exploding issue. Can LSTM solve this issue? why?
7. (5') What if we remove the forget gate? Will the model performance be increased or decreased?