IE643: Deep Learning - Theory and Practice **Jul-Dec 2023**

Scribing Assignment (Due On 11^{th} November 2023 11:59 PM IST)

1 Instructions

Answer all questions. Write your answers clearly. For Question 1, write the derivations clearly and prepare a single pdf file containing your solutions. For each exercise problem where programming is required, create a separate python notebook and make sure that all your answers are present in that notebook.

Name your python notebook files as `IE643_YOURROLLNO_scribingassignment_Ex2.ipynb`,

`IE643_YOURROLLNO_scribingassignment_Ex3.ipynb`.

Construct a folder named `IE643_YOURROLLNO_scribingassignment` which contains all the pdf and notebook files. Upload on moodle, a single zip file named `IE643_YOURROLLNO_scribingassignment.zip`, containing the folder with all the pdf and notebook files.

IMPORTANT NOTE: Make sure that your code can be run in Google colab as it is. Include the installation code for all relevant packages used in your code. No further installations should need to be performed by the TAs during evaluation.

You can score a maximum of 60 marks in this assignment.

If you answer Question 3, the marks you score in Question 3 will be considered for grade improvement.

There will be no extensions to the submission deadline.

2 Questions

1. Consider a LSTM network with an input layer which accepts d dimensional input at every time step. Consider one hidden layer which contains one lstm cell. Consider an output layer which contains K neurons. Assume that the network is trained on inputs of the form $(\mathbf{x}, \mathbf{y})$ where $\mathbf{x} = (x_1, x_2, \ldots, x_T)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_T)$. Assume a squared error loss function. Answer the following:

   (a) [**6 marks**] Discuss the expressions for forward pass for quantities $h(t)$, $C(t)$ and $\hat{y}(t)$. Using these expressions construct the expressions for squared error $e_j^2(t)$ at j-th neuron at the output layer at time t. Also write the expressions for the total error $e^2(t)$ at time t and net error E aggregated over time steps $t = 1, 2, \ldots, T$.

   (b) [**20 marks**] Using the discussions in class for RNNs, find the differential $\partial E/\partial w_{ij}$ for the following cases for the LSTM: (check LSTM lecture slides for understanding notations)

      i. the weight w_{ij} connects a neuron from hidden state $h(t-1)$ to the input gate $i(t)$
      ii. the weight w_{ij} connects a neuron from hidden state $h(t-1)$ to the forget gate $f(t)$
      iii. the weight w_{ij} connects a neuron from hidden state $h(t-1)$ to the state $\tilde{C}(t)$
      iv. the weight w_{ij} connects a neuron from hidden state $h(t-1)$ to the output gate $o(t)$

(c) [**8 marks**] Using your expressions for differentials in part (b), discuss if you get recurrence relations between $\partial E/\partial C(t)$ and $\partial E/\partial C(t+1)$. Also discuss if you get recurrence relations between $\partial E/\partial h(t)$ and $\partial E/\partial h(t+1)$.

(d) [**6 marks**] Based on your answer to part (c), discuss if the LSTM architecture is useful to preserve gradients to longer time steps when compared to RNNs.

2. [**20 marks**] Consider the colab notebook posted for Scribing Assignment Part II and the related data file. Complete the solutions for questions posted in the notebook.

3. [**Optional, only for improving grade points**] [**30 marks**] Consider the MLP code posted for Assignment 1. Adapt the code and implement the forward pass and backpropagation expressions for RNN. Assume an input layer with 28 neurons, two hidden layers with 100 neurons each with ReLU activations. Assume that the RNN works for 28 time steps. Assume an output layer only at the 28-th time step and let the output layer consist of 4 neurons with linear activations.

   Consider the MNIST dataset (both train and test) for classes $0, 5, 6, 9$. Recall that MNIST data consists of 28×28 size images. Assume that this image is passed into the RNN in the following manner: the first row of 28 pixels is input at the first time step, the second row of 28 pixels is input at the second time step and so on upto the last row of 28 pixels which is input at the 28-th time step. Consider a softmax layer after the output layer and cross-entropy loss function for the output layer.

   Choose a batch size of 32 or 64. Choose SGD optimization algorithm with learning rate tuned from the set $\{0.01, 0.001, 0.0001\}$.

   Train your RNN on the training part with appropriate validation procedures, record the train and validation losses and accuracies. Plot the losses and accuracies for train and validation data sets. Test the performance of your RNN on the test data and print the accuracy.

END OF SCRIBING ASSIGNMENT QUESTIONS