LNMIIT/B.Tech./CSE/PE/2019-20/ODD/DL/ET

LNMIIT
The LNM Institute of Information Technology

The LNM Institute of Information Technology

Computer Science and Engineering

Deep Learning (CSE4121)

Time: 3 hours | End Term | Max. Marks: 50

Date: December 04, 2019

Read the following instructions carefully:

- *There are 6 questions printed on both sides of the paper.*
- **No marks** *for providing just expressions/answers unless accompanied with correct justification and/or derivation.*
- *In case of any doubt, make your assumption, write it clearly and continue.*
- *Unnecessary text will attract negative marking. So think before you write.*

1. Assume a huge text corpus whose vocabulary size is 10,000 words.
   (a) How will you represent each word using one hot vectors?
   (b) How will the representation change if the corpus has 2,00,000 words in the vocabulary?
   (c) State two prominent disadvantages of one hot vector representation.
   (d) What are word embeddings?
   (e) How do word embeddings solve the problems mentioned by you in part (c)?

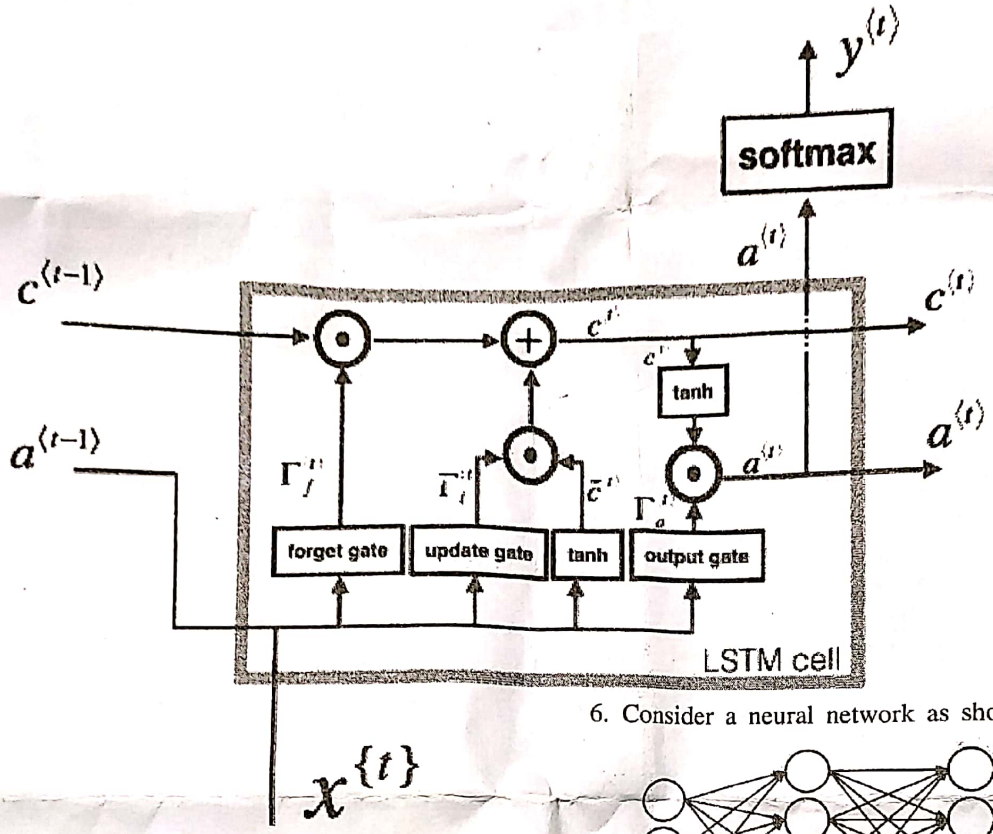
1+1+2+2+2 Marks

2. (a) Plain RNN unit is almost never used in real applications. What is actually used is gated recurrent units like GRUs and LSTMs. Do you agree with these statements? State reasons for your answer.
   (b) Draw bi-directional LSTM network.
   (c) Provide two examples where bi-directional LSTMs will be more effective than plain RNN units.
   (d) How RNNs help making a language model?
   (e) Using LSTM cell as shown, write equations for GRU.

2+2+2+2+3 Marks

3. (a) Prove that dropout regularization behaves like L2 regularization.
   (b) How to recognize overfitting?
   (c) Draw a skip connection of Resnet and write the corresponding equation for forward propagation.
   (d) Prove that Resnets have a regularization effect.

3+2+(1+2)+2 Marks



4. Assume your neural network has only two parameters w_1 and w_2.

   (a) Write down gradient descent algorithm (no marks for this) and show how the change made by virtue of using moving averages improves gradient descent.

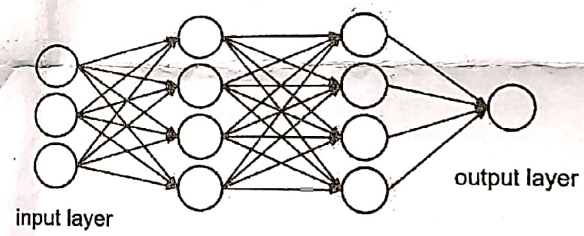
   (b) How can we use a better learning rate?

   4+2 Marks

5. You are provided a previous layer activation volume with dimensions $20 \times 20 \times 100$. Your goal is to convert this volume to another volume of size $20 \times 20 \times 200$ using operations such that the number of parameters is minimum and all the parameters in a filter are distinct. Also provide the count of parameters.

   3+1 Marks

6. Consider a neural network as shown in the figure.



   (a) Provide the dimensions of parameter matrices learnt.

   (b) Show that initializing all the parameters to zero is a bad idea.

   (c) Redraw the network with minimum changes required, if it is designed to classify three distinct classes.

   (d) Show that the softmax activation computes a probability distribution in part (c).

   (e) Assume that as a result of sparse connections and parameter sharing, only 3 parameters are being used in $W^{[2]}$. Redraw this layer clearly showing the parameters and outputs.

   2+2+2+2+3 Marks