



JK Lakshmipat University, Jaipur
Department of Computer Science and Engineering
Deep Learning (CS-1218)
End-Term, Date: May 06, 2025

Time: 2:30 hours

Semester VI B.Tech. CSE

Max marks: 70

Please read the following instructions carefully.

- There are 4 questions. The paper is printed on both sides.
- **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. **Do not call your instructor.**

1. (10*3 points) Answer the following in brief. (CO1, CO2, CO4, CO5)

- (a) Compute the number of trainable parameters in a neural network that has 10 dimensional input with 5 and 1 neurons in the hidden and output layer respectively.
- (b) Compute the gradients of $J(w_1, w_2, b) = w_1^2 + w_2^2 + b$. What does the gradient represent at the point $(2.5, 3.45, -1.23)$?
- (c) You have a language model that generates character by character, and another one generates word by word. Write one strength and one weakness of each model. NOTE: The second is word by word, not broken into tokens.
- (d) Draw a neural network corresponding to the following code and compute the number of parameters.

```
model = tf.keras.models.Sequential([  
    tf.keras.layers.Conv2D(30, (1,1), input_shape=(20, 20, 1)),  
    tf.keras.layers.MaxPooling2D(2, 2),  
    tf.keras.layers.Flatten(),  
    tf.keras.layers.Dense(10, activation='softmax')  
])
```

- (e) You have an input image of size 3×3 pixels and a convolution filter of size 2×2 . Show how parameter sharing and sparsity of connections occur compared with a fully connected neural network.
- (f) The idea behind the GloVe word embeddings is to use the global properties of what was only locally exploited in word2vec. Which global statistics are being talked about and how is it exploited by GloVe?
- (g) Why is non linear operation necessary in every layer of a feed forward neural network? Show what may happen in their absense.
- (h) You have a neural network as provided in part (a) in which dropout is being used. Draw one probable neural network that may be trained if dropout probability = 0.2 is used. How many such subnetworks are possible in total?
- (i) In YOLO, assume 3 neighboring grids have detected the same object, i.e., object is one but three grid cells are claiming it. How does YOLO handle such a situation?

- (j) How will you use Siamese network to detect whether the two input images belong to the same person or not?
2. Assume a recurrent neural network with input size = 3 and hidden state size = 2. The RNN is being used to do named entity recognition, i.e., it is just expected to classify whether a word in a named entity or not, no further classification is needed. (CO2, CO4)
- (2 points) Draw a plain recurrent neural network to accomplish this task.
 - (3 points) Provide the dimensions of all the parameter matrices involved in this neural network.
 - (3 points) Show that the network suffers from vanishing or exploding gradients problem.
 - (2 points) If we convert this RNN to a LSTM with the gates of the LSTM having the same structure as the above RNN, how many parameters will that LSTM have?
3. Consider word2vec algorithm. (CO1, CO2, CO3)
- (2 points) Prepare training dataset for word2vec, if the corpus is "I ate ice cream".
 - (2+1 points) Draw a neural network corresponding to word2vec if corpus is as defined above and the solicited word embedding size is 3. Find the number of parameters in the network.
 - (3 points) Write down the cost function for it without negative sampling. Explain the terms.
 - (1+1 points) What is the issue with the above cost function and how does negative sampling solve it?
4. Recall the transformer architecture as presented in "Attention is all you need". Assuming that the total number of tokens in the shared vocabulary is 37,000, each vector is of size 512, the number of heads is provided to be 8. Feel free to use any other missing hyperparameter if needed. (CO1, CO2, CO4)
- (2 points) Provide the dimensions of the parameter matrices in the input layer.
 - (3 points) Provide the dimensions of the parameter matrices in a single self attention layer with 8 heads.
 - (2+2 points) Provide the dimensions of the parameters in computation of the cross attention with due emphasis on where the inputs to this layer are coming from.
 - ((2+1)+(2+1) points) Consider a toy example sentence, "The Sun rises in the east". How will you compute self attention for this example? Make a suitable diagram to show all the computations of self attention. How will you compute masked self attention for this example? Draw a suitable diagram. No need to assume multihead, just do it for one head.
 - (2 points) In the absense of recurrent connections, the input to the transformers is not a sequence of words anymore but just a bag of words. How does the transformer solve this problem?
 - (1+2 points) Suppose you are given a dataset of 50,000 movie reviews. You are supposed to determine the sentiment of the review at a scale of 1 to 5. You take a pretrained transformer that is trained on huge amount of text but only on the six encoder layers of the transformer as provided above. Using this network of six encoders stacked on each other as the base, design a neural network to solve the sentiment classification problem. Reason for your network.