



JK Lakshmipat University, Jaipur
Department of Computer Science and Engineering
Deep Learning (CS1218)
End-Term, Date: December 12, 2025
B.Tech. Semester V CSE+AI (Core), Elective

Time: 2:30 hours

Max mark: 70

Roll Number: _____

Please read the following instructions carefully.

- There are 7 questions 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. (2*4+2 points) Consider a transformer model for which d_{model} is fixed to be 512, the number of layers $l = 6$, the number of heads $n_h = 8$, and the feedforward neural network has $d_{ffnn} = 4 \times 512$ then,
- (a) Compute the number of parameters in this transformer model. You may ignore the number of parameters in layernorm, positional embeddings, word embeddings at both input and output. Thus focus mainly on the encoder-decoder of the transformer blocks.
 - (b) Draw the diagram of the transformer.
- (CO1, CO3)
2. ((2+2)+(2+2)+2+(2+2) points) You are supposed to use the attention mechanism in an encoder-decoder architecture. You are also given that the size of encoder and decoder hidden states are 100×1 and the word embedding sizes in both the languages is 300×1 . The translation is being done from English to Hindi. For the purpose of this question, assume that the statement being translated is "I want to drink water".
- (a) Show how will this be done through a diagram of the network with properly labeled encoder, decoder, and the required states.
 - (b) If LSTMs are used in the encoder decoder then find the number of parameters in the encoder and the decoder.
 - (c) Find the number of parameters specific to the attention network.
 - (d) Label the vectors necessary for the computation of attention scores. How will the attention scores be computed for the third time stamp of the decoder?
- (CO2, CO3, CO4)
3. (2+2+2+(2+2) points) You are training a language model using RNNs.
- (a) Draw a diagram showing the training process.
 - (b) Write down the loss function for language modeling.

- (c) Suppose you are using teacher forcing to train the language model, how will your training be impacted? Write precisely.
- (d) Mathematically argue why plain RNNs suffer from vanishing and exploding gradient problems.

(CO2, CO3)

4. (2+2+(2+2)+2 points) You are training YOLO for object detection. Assume that the input images are divided into 10×10 grids (in the class it was 19×19). There are 3 object classes and a background.

- (a) If 4 adjacent cells have detected the same object how will the network reduce the number of output bounding boxes?
- (b) State the situations in which multiple grid cells may claim the same object in YOLO.
- (c) If 2 anchor boxes are used, how will the labels for each input image will be decided?
- (d) Mr. K suggests to use 5 anchor boxes. Argue in favour or against his choice.

(CO1, CO2, CO4)

5. (2+(2+2) points) You are provided two images containing two faces which may or may not be of the same person. Your goal is to detect if both the pictures belong to the same person.

- (a) It is a binary classification problem. Formulate it so as to solve the above problem.
- (b) Write the triplet loss and argue why it is important?

(CO2, CO3)

6. ((2+2+2)+2+2 points) (a) Why is Mean Squared Error used as a cost function for linear regression and cross-entropy for logistic regression? Write their expressions. Label them well.

- (b) Assuming four input features are (x_1, x_2, x_3, x_4) , draw a linear regression network and a logistic regression network. Neatly label all the elements of both the networks along with their input and output.

- (c) Prove that logistic regression is limited to linear classifiers only.

(CO2, CO3, CO4)

7. ((2+2)+3+3 points) (a) Show that CNNs are feed forward neural networks with weight sharing and sparse parameters.

- (b) What is the similarity between the ResNets and LSTMs?
- (c) Why is deep learning suddenly so successful in the last decade whereas the neural networks have been around for almost a century now? Mention three strong reasons.

(CO2, CO4)