Indian Institute of Technology Patna
End Semester Examination
Course Name: Large Language Models
Course Code: MA6119
Date: 30-11-2024

Maximum Marks: 60 **Time: 3 Hours**

---

1. **Fully Connected Neural Networks** **[Total Marks = 13]**

(a) Given a text classification task with 10,000 unique words in the vocabulary, you design a neural network with the following fully connected layers:

- **Input layer**: 300-dimensional word embeddings.
- **Hidden layer 1**: 128 units, ReLU activation.
- **Hidden layer 2**: 64 units, ReLU activation.
- **Output layer**: 5 units (for 5 classes), softmax activation.

If the model uses a fully connected architecture, calculate the total number of trainable parameters (weights and biases) in the network. **[Marks = 2]**

(b) How does splitting a dataset into train, val and test sets help identify underfitting? **[Marks = 2]**

(c) You are designing a deep learning system to detect hate speech in social media posts. It is crucial that your model identifies as much hate speech as possible to prevent harmful content from spreading. Which of the following is the most appropriate evaluation metric: Accuracy, Precision, Recall, Loss Value? Explain your choice. **[Marks = 2]**

(d) You have a single hidden-layer neural network for a binary text classification task, where the input is $X \in \mathbb{R}^{n \times m}$ representing a bag-of-words (BoW) or word embeddings for m input samples, and the output $\hat{y} \in \mathbb{R}^{1 \times m}$ represents the predicted class probabilities for each sample. The true label is $y \in \mathbb{R}^{1 \times m}$. The forward propagation equations are as follows:

$$z^{[1]} = W^{[1]}X + b^{[1]}$$
$$a^{[1]} = \sigma(z^{[1]})$$
$$\hat{y} = a^{[1]}$$

The cost function is:

$$J = -\sum_{i=1}^{m} \left[y^{(i)} \log(\hat{y}^{[i]}) + (1 - y^{(i)}) \log(1 - \hat{y}^{[i]}) \right]$$

Write the expression for $\frac{\partial J}{\partial W^{[1]}}$ as a matrix product of two terms. **[Marks = 4]**

(e) Consider a neural network encoder $z = \text{softmax}[f_\theta(X)]$. You can think of f_θ as an MLP for this example. z is the softmax output, and we want to discretize this output into a one-hot representation before passing it into the next layer.

Consider the operation one_hot, where one_hot(z) returns a one-hot vector with a 1 at the argmax location. For example:

$$\text{one\_hot}([0.1, 0.5, 0.4]) = [0, 1, 0].$$

Say we want to pass this output to another fully connected layer g_ϕ to get a final output y.

(i) Is there a problem with the neural network defined below? **[Marks = 1]**

$$y = g_\phi(\text{one\_hot}(\text{softmax}(f_\theta(X))))$$

(ii) Consider the following function:

$$z = S_\tau(f_\theta(X)) = \text{softmax}\left(\frac{f_\theta(X)}{\tau}\right)$$

Here dividing by τ means every element in the vector is divided by τ. Obviously, when $\tau = 1$, this is exactly the same as the regular softmax function. What happens when $\tau \to \infty$? What happens when $\tau \to 0$?
Hint: You don't need to prove these limits; just showing a trend and justifying is sufficient. **[Marks = 2]**

2. **RNNs, LSTMs and Transformer** **[Total Marks = 21]**

(a) Consider a 3-layer RNN where:

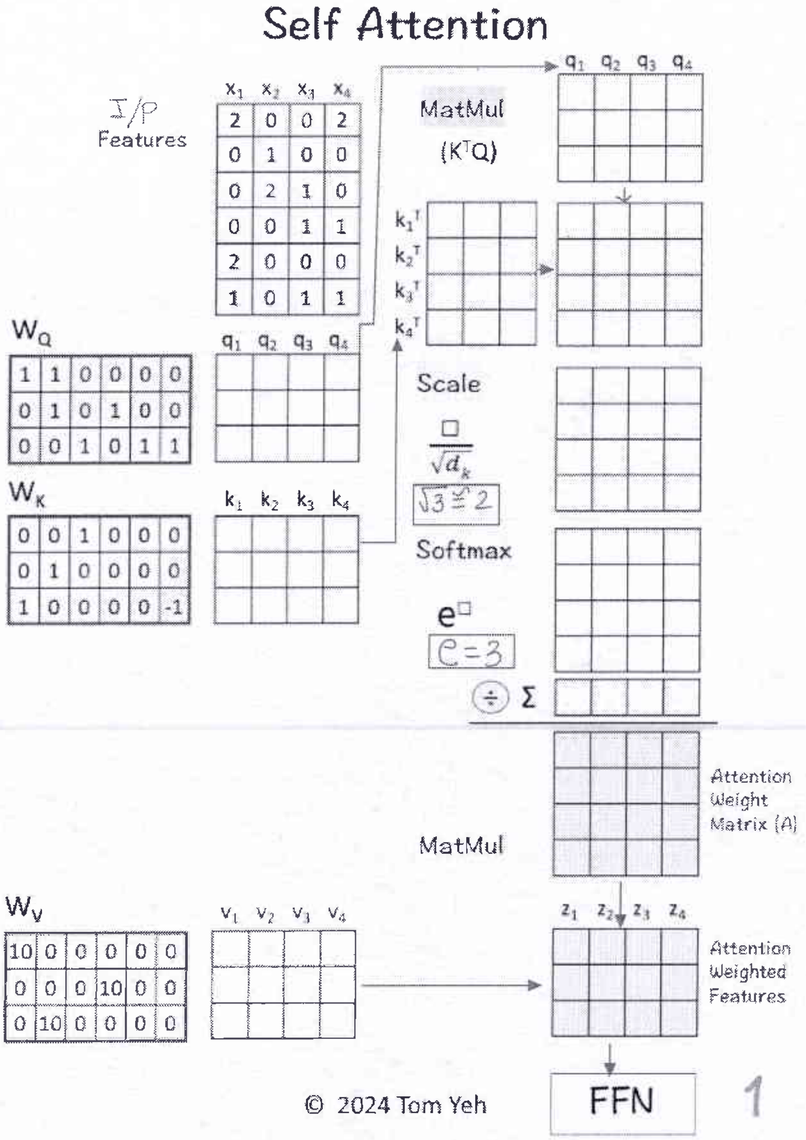
- The input dimension is 6.
- Each layer has 8 hidden units.
- The output dimension is 5.

How many trainable parameters (weights and biases) are there in total across all three layers? **[Marks = 3]**

(b) What challenges do Recurrent Neural Networks (RNNs) face during training due to the behavior of gradients, and how do these issues affect the learning process over long sequences? **[Marks = 2]**

(c) Provide the equations for a standard Long Short-Term Memory (LSTM) cell and explain the role of each gate (input gate, forget gate, output gate) and the cell state. **[Marks = 3]**

(d) Perform the right operations and fill in the blanks: **[Marks = 6]**



Self Attention

I/P Features

x_1	x_2	x_3	x_4
2	0	0	2
0	1	0	0
0	2	1	0
0	0	1	1
2	0	0	0
1	0	1	1

W_Q

1	1	0	0	0	0
0	1	0	1	0	0
0	0	1	0	1	1

W_K

0	0	1	0	0	0
0	1	0	0	0	0
1	0	0	0	0	-1

W_V

10	0	0	0	0	0
0	0	0	10	0	0
0	10	0	0	0	0

© 2024 Tom Yeh

(e) You are tasked with building a model to predict daily stock prices based on historical data. The data exhibits strong temporal dependencies, with short-term fluctuations and occasional long-term trends. The dataset is relatively small, and each sequence has about 30 time steps. Which architecture (RNN, LSTM, or Transformer) would be most suitable for this task? Justify your choice considering the size of the dataset and the temporal dependencies. **[Marks = 3]**

(f) Design a Visual Question Answering (VQA) system where the input consists of an image and a natural language question, and the output is a text-based answer. Describe how you would extract features from the image and process the text input. How would you fuse the image and text features to generate the final answer? **[Marks = 4]**

3. **Word2Vec, BERT and GPT** **[Total Marks = 15]**

(a) How does Word2Vec differ from traditional one-hot encoding for word representation? **[Marks = 2]**

(b) What is the primary training objective of the Word2Vec model? How does it generate word embeddings? **[Marks = 2]**

(c) Explain the two main pre-training tasks of BERT: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). How do these tasks help in learning contextual representations? **[Marks: 3]**

(d) You are tasked with developing a sentiment analysis system for a product review platform to classify reviews as positive, negative, or neutral. Given the need to capture contextual meanings and handle nuanced language, you decide to use BERT. Explain how you would preprocess the review data for BERT, fine-tune the model for sentiment classification, and use it to predict sentiments for new reviews. Additionally, discuss the advantages of BERT in this context. **[Marks: 3]**

(e) What is GPT, and how does it generate human-like text based on the input it receives? **[Marks: 3]**

(f) Provide one use case where GPT outperforms BERT and explain why? **[Marks: 2]**

4. **ChatGPT and RAG** **[Total Marks = 11]**

(a) Explain how ChatGPT generates responses and describe the role of Reinforcement Learning with Human Feedback (RLHF) in fine-tuning the model. Compare RLHF with traditional supervised learning, using an example to highlight the differences in their training approaches. **[Marks: 5]**

(b) Explain the key components of the RAG pipeline and provide an example use case where it would be particularly effective. **[Marks: 4]**

(c) In RAG, if the retriever component fails to find relevant documents, what impact will this have on the generator's output? **[Marks = 2]**