Answeer Key

Multiple choice questions. 1 marks for the right answer, -2 marks for the wrong answer and -1 marks for not attempting the question. All the questions are based on the Attention is all you need paper.

1. What is a primary advantage of the Transformer architecture over RNN-based models?
   a. It eliminates the need for recurrent connections.
   b. It allows for faster training by enabling parallelization.
   c. It requires significantly fewer GPUs for training.
   d. It uses convolutional layers to improve accuracy.
2. What role does multi-head attention play in the Transformer model?
   a. It enables the model to attend to different parts of the input simultaneously.
   b. It allows the model to represent different relationships in the data.
   c. It enhances memory retention within the model.
   d. It improves the interpretability of the model.
3. Which of the following are components of the encoder stack in the Transformer model?
   a. Multi-head self-attention.
   b. The attention dot product is scaled to restrict its variance before applying softmax.
   c. Position-wise feed-forward networks.
   d. Dropout layers.
4. Given an input of sequence length "n" and embedding dimension "d" (nXd). Derive the time complexity of the self attention layer of the encoder block, given the dimension of values to be twice that of the query and keys. How is it different from the recurrent and convolution layers' time complexity? Explain any new variables or assumptions you introduce in the answer. Attention is all you need Table 1. [7 marks]

input = (N, d)

↓

query = (N, d)

key = (N, d)

values = $(N, 2d)$

attention matrix = (N, N) → (Right answer)

time complexity = $(N^2 d)$

Rescaling values = $(N, 2d)$

time complexity = N^2

total time complexity = $O(N^2 d + N^2)$

$= O(N^2 \cdot d)$

---

for the correct answer: 5 marks

without explanations = 2 marks.

Recurrent = $O(n \cdot d^2)$

convolution = $O(k \cdot n \cdot d^2)$

for the correct answer = 2 marks

without explanation = 1 marks.

→ Scales with d^2

→ Scales with d^2 but no seq operation