Assignment 3



Q1 10 Points

Q1.1 BERT 5 Points

What is the optimization objective of BERT?

Predicting a masked word.

Predicting whether two sentences follow each other.

Both a and b.

Explanation:



Because we replaced recurrent connections with attention modules.

Because it decreases overfitting in RNN and transformers.

Explanation:				
	\neg			

Q2 Time Complexity of Transformers 10 Points

Q2.1 5 Points

What is the time complexity of Transformers as a function of the number of heads h?

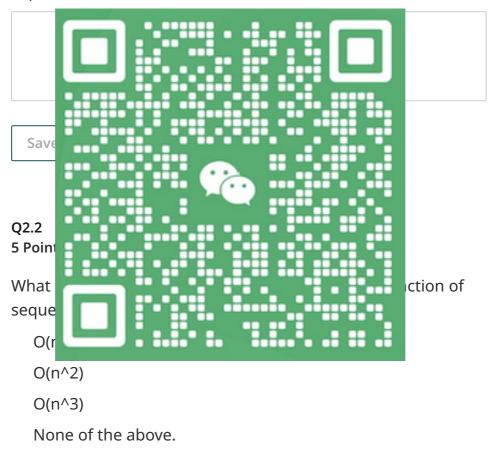
O(log h)

O(h)

O(h^2)

None of the above.

Explanation:



Explanation:

Q3 Transformers 10 Points

Q3.1 5 Points

What is the main difference between the Transformer and the encoder-decoder architecture?

- Unlike encoder-decoder, transformer uses attention only in encoders.
- Unlike encoder-decoder, transformer uses attention only in decoders.
- Unlike encoder-decoder, transformers have multiple encoder-decoder structures layered up together.



Q3.2 5 Points

Which of the following architectures cannot be parallelized?

CNNs

RNNs

Transformers

Explanation:



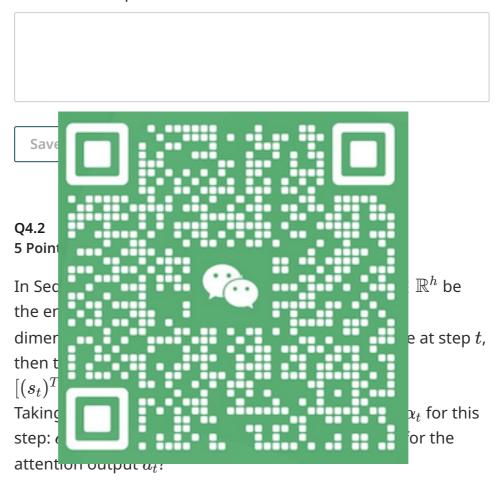


Q4 Transformers 10 Points

Q4.1 5 Points

For the vectors x_i , consider the weighted average $y_i = \sum_j \alpha_{i,j} \cdot x_j$ where $w_{i,j} = x_i^T x_j$ and $\alpha_{i,j} = \operatorname{softmax}(w_{i,j})$. What is $\sum_j \alpha_{i,j}$ for any i?

Answer and Explanation:



Please type any math input using latex format enclosed by \$\$...\$\$.

Answer and Explanation:

Q5 GNNs 10 Points

Q5.1 5 Points

What are the two key operations used for updating a node representation in a GNN?

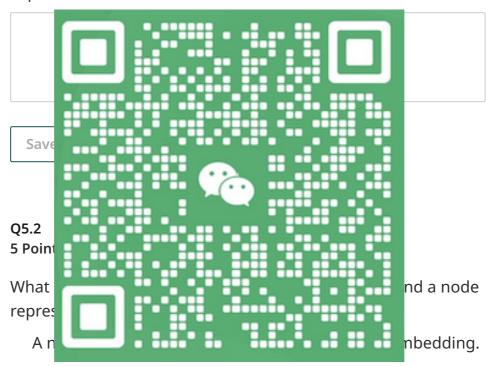
Aggregate and Combine.

Aggregate and Message.

Combine and Update.

Aggregate and Max Pooling.

Explanation:



A node embedding is a special case of node representation.

There is no difference between the two.

Exp	lanation:			
Sa	ave Answer			

Q6 GNNs 10 Points

Q6.1 5 Points

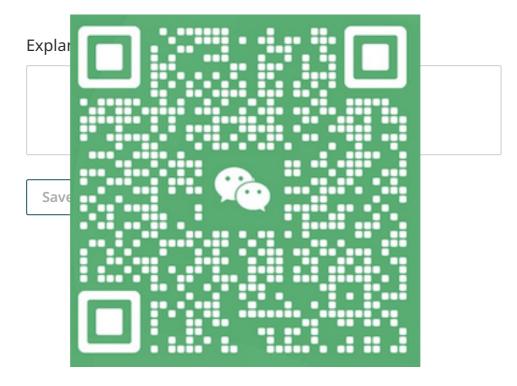
Citation networks can be treated as graphs where researchers cite fellow researchers' papers. Given a paper "A" we would like to predict if the paper cites another paper "B" or not. Which type of prediction task can you model this to be?

Node prediction.

Link prediction.

Graph prediction.

Sub graph prediction.



Q6.2 5 Points

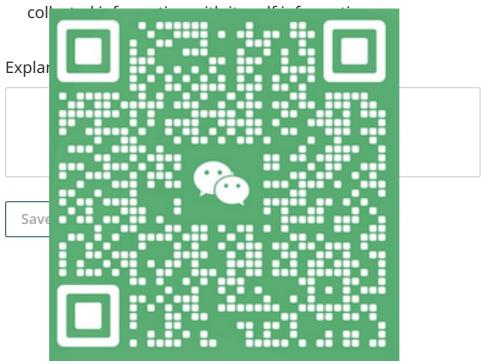
Select the correct statement among the following:

Combine operation gathers information from all nodes and aggregate operation updates the collected information with its self information.

Aggregate operation gathers information from all nodes and combine operation updates the collected information with its self information.

Combine operation gathers information from it's neighboring nodes and aggregate operation updates the collected information with its self information.

Aggregate operation gathers information from it's neighboring nodes and combine operation updates the



Q7 GNNs

10 Points

What is the Laplacian matrix for a graph with nodes $\{1,2,3,4,5\}$ and edges $\{(1.5),(1,3),(2,3),(2,5),(3,4)\}$?

Answer and Explanation:
You can also upload a picture of your work: Please select file(s) Select file(s) Q8 RN 10 Poil Given $ \begin{pmatrix} -1, \\ 0, -\\ (x_1, x_1) \end{pmatrix} $ $=$
Answer Explanation:
You can also upload a picture of your work: Please select file(s) Save Answer

Q9 Convolutional view of a linear RNN 10 Points



Q9.1 8 Points

Given an RNN defined by

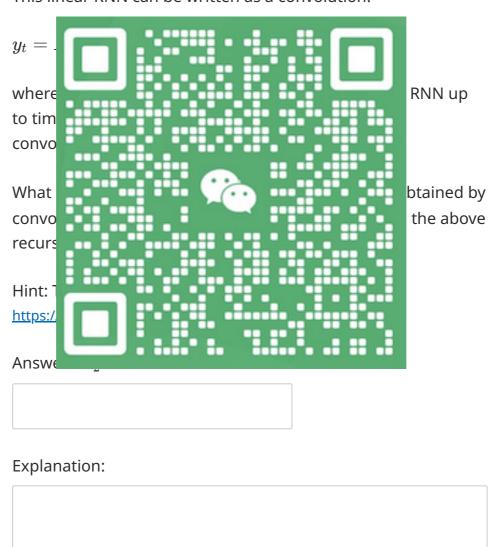
$$s_t = W \cdot s_{t-1} + U \cdot x_t$$

 $y_t = C \cdot s_t$
 $s_0 = U \cdot x_0$

where $s_t \in \mathbb{R}^2$, $x_t \in \mathbb{R}$ and $y_t \in \mathbb{R}$ denote the hidden state, input, and output of the RNN at timestep t, respectively.

$$\mathsf{W} = \begin{pmatrix} -1, 0 \\ 0, -1 \end{pmatrix}, \, \mathsf{U} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \, \mathsf{C} = \, \begin{pmatrix} -1, 1 \end{pmatrix}$$

This linear RNN can be written as a convolution:



You can also upload a picture of your work:

Please select file(s)

Select file(s)

Save Answer

Q9.2 2 Points

Give one reason why we might want to implement a linear RNN using a convolution instead of recursion:



Q10 Parameter Efficient Tuning of Attention layers 10 Points

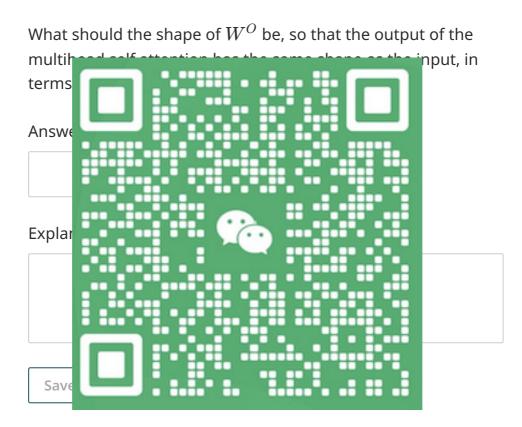
Q10.1 2.5 Points

Consider a multihead self attention block with h heads.

$$MultiHead(X) = Concat(head_1, ..., head_h)W^O$$

where
$$head_i = \operatorname{Attention}(XW_i^Q, XW_i^K, XW_i^V)$$
.

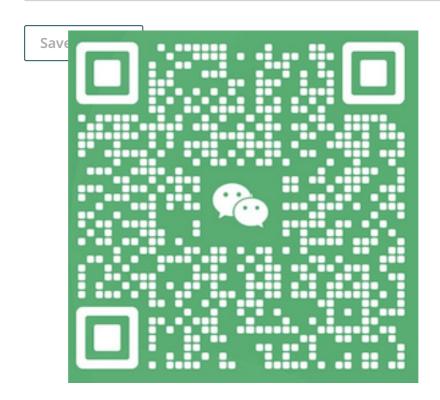
Assume bias=False,
$$X \in \mathbb{R}^{N imes D}$$
, $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{D imes d}$.



Q10.2 2.5 Points

How many parameters are there in total, in terms of h, d, D, and N?

Answer:		
Explanation:		



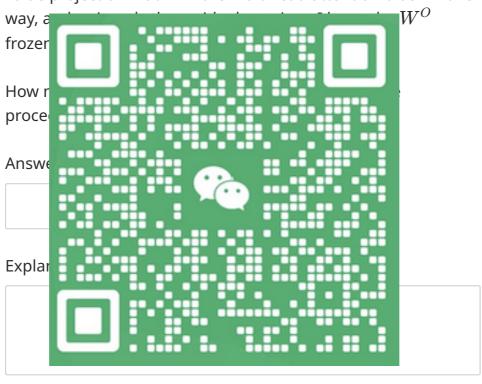
Q10.3 2.5 Points

I have some data I want to finetune the multihead attention block on, but I don't want to finetune all the parameters. Specifically, I want to use LoRA to finetune the model. (LoRA is described here: https://arxiv.org/pdf/2106.09685)

Specifically, for every trainable matrix W, I decompose it as:

$$W = W_0 + \Delta W = W_0 + BA$$

 W_0 is the pretrained weights that are not trained. $\Delta W = BA$ is a residual matrix with the same shape as W that I do train. $B \in \mathbb{R}^{D \times r}$ and $A \in \mathbb{R}^{r \times d}$. I decompose every query, key and value projection matrix in the multihead attention block in this



Q10.4 2.5 Points

State two separate advantages of finetunin	g a model with
LoRA:	

Save Answer

Save All Answers

Submit & View Submission >

