1

Question: What is the name of the novel network architecture proposed in the paper?

A) Recurrent Neural Network

B) Convolutional Neural Network

C) Transformer

D) Encoder-Decoder Network

Correct Answer: C) Transformer

2

Question: What mechanism is the Transformer solely based on?

A) Recurrence

B) Convolution

C) Attention

D) Residual connections

Correct Answer: C) Attention

3

Question: What is the main advantage of the Transformer over recurrent models according to the paper?

A) Higher accuracy on sentiment analysis

B) Improved performance on object detection

C) More parallelization during training

D) Reduced memory consumption during inference

Correct Answer: C) More parallelization during training

4

Question: What is the dimensionality (d_model) of the output from the embedding layers in the Transformer?

A) 64

B) 128

C) 256

D) 512

Correct Answer: D) 512

5

Question: What type of attention mechanism is used to relate different positions of a single sequence?

A) Self-attention

B) Multi-head attention

C) Encoder-decoder attention

D) Additive attention

Correct Answer: A) Self-attention

6

Question: How many parallel attention layers (heads) are used in the Multi-Head Attention mechanism?

A) 2

B) 4

C) 8

D) 16

Correct Answer: C) 8

7

Question: What activation function is used in the position-wise feed-forward network?

A) Sigmoid

B) Tanh

C) ReLU

D) Linear

Correct Answer: C) ReLU

8

Question: What technique is used to inject information about the order of tokens in the sequence?

A) Positional encoding

B) Recurrent connections

C) Convolutional filters

D) Bag-of-words representation

Correct Answer: A) Positional encoding

9

Question: What type of optimizer is used to train the Transformer model?

A) Stochastic Gradient Descent (SGD)

B) Root Mean Square Propagation (RMSprop)

C) Adam

D) Adagrad

Correct Answer: C) Adam

10

Question: What regularization technique is used to prevent overfitting in the Transformer model?

A) L1 regularization

B) L2 regularization

C) Dropout

D) Early stopping

Correct Answer: C) Dropout

11

Question: What is the BLEU score achieved by the Transformer (big) model on the WMT 2014 English-to-German translation task?

A) 24.6

B) 26.3

C) 27.3

D) 28.4

Correct Answer: D) 28.4

12

Question: What dataset was used for evaluating the Transformer's performance on English constituency parsing?

A) Universal Dependencies (UD)

B) Penn Treebank

C) Brown Corpus

D) GLUE benchmark

Correct Answer: B) Penn Treebank

13

Question: What is the maximum path length between any two input and output positions in a self-attention layer?

A) O(1)

B) O(log(n))

C) O(n)

D) O(n/k)

Correct Answer: A) O(1)

14

Question: What is the purpose of the scaling factor in the Scaled Dot-Product Attention mechanism?

A) To normalize the attention weights

B) To prevent the dot products from growing too large

C) To improve the computational efficiency

D) To handle variable-length sequences

Correct Answer: B) To prevent the dot products from growing too large

15

Question: Which of the following is NOT a type of attention used in the Transformer?

A) Encoder-decoder attention

B) Self-attention in the encoder

C) Cross-attention in the decoder

D) Self-attention in the decoder

Correct Answer: C) Cross-attention in the decoder

16

Question: What byte-pair encoding vocabulary size was used for the English-German translation task?

A) 16,000

B) 25,000

C) 32,000

D) 37,000

Correct Answer: D) 37,000

17

Question: What is the purpose of masking in the decoder's self-attention layer?

A) To prevent attending to future positions

B) To focus on specific parts of the input sequence

C) To reduce the computational complexity

D) To handle padding tokens

Correct Answer: A) To prevent attending to future positions

18

Question: What is the purpose of the residual connections used in the Transformer?

A) To improve the flow of gradients during training

B) To reduce the number of parameters in the model

C) To prevent overfitting

D) To handle variable-length sequences

Correct Answer: A) To improve the flow of gradients during training

19

Question: According to the paper, what type of attention generally performs similarly to dot-product attention for small values of d_k?

A) Self-attention

B) Multi-head attention

C) Additive attention

D) Encoder-decoder attention

Correct Answer: C) Additive attention

20

Question: What is the approximate training time for the Transformer base model on eight P100 GPUs?

A) 12 hours

B) 3.5 days

C) 1 week

D) 2 weeks

Correct Answer: A) 12 hours