CS379-Machine Learning

Marks: 20 Marks

Name:__________ **Roll no.:**__________ **Date:**__________

Instructions: Questions 1–14 carry 1 mark each and Questions 15–17 carry 2 marks each. All answers must be written on the back-side of the page. Best of luck!

Q1. Multi-head attention in Transformers allows:

A. Fewer computations

B. Multiple representations of data

C. Faster convergence

D. Better noise handling.

Q2. The attention mechanism in Transformers helps:

A. Avoid vanishing gradients

B. Reduce parameters

C. Provide global context

D. Improve dropout regularization

Q3. What makes an LSTM able to learn long-range dependencies better than vanilla RNNs?

A. Increased hidden units

B. ReLU activation

C. Forget gate mechanism

D. Deeper network layers

Q4. Gated Recurrent Units (GRU) differ from LSTM by:

A. Using fewer gates

B. Using convolutional layers

C. Adding dropout gates

D. Removing hidden states

Q5. What problem does gradient clipping aim to solve in RNNs?

A. Overfitting

B. Vanishing gradients

C. Exploding gradients

D. Underfitting

Q6. Positional encoding in transformers is used primarily because:

A. Transformers cannot inherently encode sequence order.

B. It prevents overfitting.

C. It helps attention convergence.

D. It reduces computational complexity

Q7. A denoising autoencoder explicitly learns to:

A. Compress data more effectively.

B. Reconstruct noisy inputs into original data.

C. Classify noisy inputs into labels.

D. Generate diverse outputs

Q8. VAEs optimize their objective using:

A. Adversarial training

B. Backpropagation through expectation maximization

C. Variational inference and reconstruction loss

D. Purely supervised learning

Q9. Mode collapse in GANs refers to:

A. Generator producing outputs from a very limited subset.

B. Discriminator becoming overly accurate.

C. Training process becoming slow.

D. Generating noisy samples only.

Q10. A GAN generator learns primarily through:

A. Directly reconstructing input data.

B. Maximizing discriminator loss.

C. Supervised classification loss.

D. Minimizing discriminator's ability to classify outputs.

Q11. What is the time complexity of attention in Transformers?

A. O(1)

B. O(n)

C. O(n^2)

D. O(log n)

Q12. Why is masking necessary in Transformers for sequence modeling?

A. To prevent models from attending to future tokens during training

B. To eliminate rare words

C. To ensure equal weights to all words

D. To speed up training

Q13. What is the role of the hidden state in an RNN?

A. To store the dataset.

B. To control the weights of the network.

C. To remember the output of previous time steps and pass it to the next step.

D. To initialize the input layer.

Q14. In VAEs, what is typically learned in the bottleneck (latent) space?

A. A single deterministic point

B. A probability distribution

C. A reconstruction of the input

D. A sequence of input embeddings.

Q15. A vanilla RNN processes a sequence of 5 time steps with an input vector of size 10 and hidden state size 20. How many parameters does the hidden-to-hidden weight matrix have?

A. 100

B. 200

C. 400

D. 2000

Q16. In a GAN, if the generator takes a 100-dimensional noise vector and produces a 28×28 image, how many output units does the generator have?

A. 28

B. 100

C. 128

D. 784

Q17. What is the KL divergence between $\mathcal{N}(0,1)$ and $\mathcal{N}(0,1)$?

A. 0

B. 1

C. 0.5

D. undefined

Q1.		**Q10.**	
Q2.		**Q11.**	
Q3.		**Q12.**	
Q4.		**Q13.**	
Q5.		**Q14.**	
Q6.		**Q15.**	
Q7.		**Q16.**	
Q8.		**Q17.**	
Q9.			