Monsoon 2025 End-Semester Examination

Advances in Deep Learning (ECE/CSE 677)

Time: 1 hour Max Marks: 13

Roll No:
Name:
General Instructions:

- Questions may have more than one correct answer. Mark all correct ones.
- Q1-Q13 carry 1 mark each. Q14 caaries 0.5 marks.

Q1. Consider a standard Transformer encoder operating on a sequence of token embeddings $x_1, \ldots, x_T$. Which of the following are true about *self-attention without positional information*? [CO1]

i. It is permutation equivariant w.r.t. the sequence of tokens.

ii. It can distinguish between the sequences (x_1, x_2, x_3) and (x_3, x_2, x_1).

iii. It can still model pairwise interactions between tokens.

iv. It is invariant to any reordering of tokens.

Q2. For a Vision Transformer (ViT) with N image patches and embedding dimension d, the computational complexity of **full self-attention** in one layer (ignoring constants and number of heads) is: [CO1]

A. $\mathcal{O}(N)$

B. $\mathcal{O}(Nd)$

C. $\mathcal{O}(N^2)$

D. $\mathcal{O}(N^2 d)$

Q3. Which statements about ViT are correct? [CO3]

A. ViT replaces convolutional layers with patch embeddings followed by Transformer layers.

B. ViT follows a hierarchical type of architecture, similar to that of CNNs.

C. ViT typically requires large-scale pretraining to match or beat CNNs on image tasks.

D. ViT inherently encodes locality biases like CNNs through weight sharing over spatial neighborhoods.

Q4. In adversarial training with PGD attacks, which statements are true? [CO2]

A. Adversarial examples are generated during each iteration in training.

B. The loss is computed only on clean examples.

C. A regularizer that aligns the logits or softmax probabilities of clean and adversarial examples can improve training

D. Adversarial samples may lie off the manifold of clean data that can complicate adversarial training.

Q5. The role of **temperature** T in the softmax during distillation is to: [CO1]

A. Smooth the teacher's output distribution to reveal "dark knowledge" about non-argmax classes.

B. Directly control the learning rate of the student optimizer.

C. Reduce the dimensionality of logits.

D. None of the above.

Q6. When training the student in distillation: [CO3]

A. A higher temperature is often used in the soft loss, then the same T is used at test time.

B. The student is typically trained using a combination of soft targets from the teacher and true labels.

C. Distillation is only valid if student and teacher have identical architectures.

D. Distillation can be used to compress very large models into much smaller ones.

Q7. Consider Model-Agnostic Meta-Learning (MAML) in a few-shot classification scenario. The meta-objective of MAML is to: [CO2]

A. Learn parameters that can perform well for unseen classes.

B. Learn an initialization θ that can be adapted to many tasks with few gradient steps.

C. Learn a separate optimizer that replaces gradient descent.

D. None of the above.

Q8. In a 1-shot-N-way classification episode under MAML, which of the following steps are performed **during meta-training**? [CO2]

A. Sample a task and its support + query sets.

B. Perform inner-loop updates on support data to get task-adapted parameters.

C. Compute meta-loss on query set using task-adapted parameters.

D. Update initial parameters θ using gradients from meta-loss.

Q9.In score-based generative modeling: [CO1]

A. The score function is defined as $\nabla_x \log p(x)$.

B. Langevin dynamics uses the score to gradually move samples toward high-density regions.

C. The score is only defined for Gaussian distributions.

D. Annealed Langevin dynamics uses a sequence of noise scales during sampling.

Q10. A standard GCN layer with input features $H^{(l)}$ and output $H^{(l+1)}$ is given by (single correct): [CO3]

A. $H^{(l+1)} = \sigma(AH^{(l)}\Theta^{(l)})$

B. $H^{(l+1)} = \sigma\big(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^{(l)}\Theta^{(l)}\big)$

C. $H^{(l+1)} = \sigma\big(LH^{(l)}\Theta^{(l)}\big)$

D. $H^{(l+1)} = \Theta^{(l)}H^{(l)}$

Q11. The Inception Score (IS) uses a pretrained classifier and is high when: [CO1]

A. $p(y \mid x)$ has low entropy for generated images.

B. The marginal $p(y)$ over generated images has high entropy.

C. Generated images are diverse across classes.

D. Generated images are pixel-wise close to real images from the training set.

Q12. The Fréchet Inception Distance (FID) between real and generated images: [CO1]

A. Compares distributions of features in an embedding space via Gaussian approximation.

B. Is invariant to any invertible linear transformation of the feature space.

C. Requires paired real and generated images (same content) to compute.

D. Is sensitive to both diversity and quality of generated samples.

Q13. Let L be the (normalized) graph Laplacian and $U\Lambda U^\top$ its eigendecomposition. Which statements are correct? [CO3]

A. Spectral graph convolution can be defined as $g_\theta * x = U g_\theta U^\top x$.

B. Vanilla spectral convolution requires eigen-decomposition of L.

C. ChebNet uses Chebyshev polynomial approximation of $g_\theta(\Lambda)$.

D. GCN (Kipf & Welling) can be seen as a first-order approximation of ChebNet.

Q14. State True of False. [CO3]

In a standard GCN layer, node features are updated using only the adjacency matrix A without any degree normalization or self-loops.