UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

CS 498
Part I Exam

02/18/2026

Name: ____

NetID: ____

This exam contains 9 pages (including this cover page) and 30 questions, The point total is 100.

All questions in this exam are single- or multiple-choice questions. Please clearly circle your answer. If you change your response, erase your previous markings cleanly. Unless otherwise specified, each question has one correct answer. If we cannot find/cannot read your solution, we reserve the right to give that question/subquestion a 0.

Distribution of Marks

Question:	1	2	3	4	5	6	7	8
Points:	2	2	2	2	2	2	2	2
Score:								

Question:	9	10	11	12	13	14	15	16
Points:	2	2	4	4	4	4	4	4
Score:								

Question:	17	18	19	20	21	22	23	24
Points:	4	4	4	4	4	4	4	4
Score:								

Question:	25	26	27	28	29	30		Total
Points:	4	4	4	4	4	4		100
Score:								

Part 1: True/False

Please clearly circle your answer. Each question is worth 2 points. If you change your answer, please erase your previous marking cleanly.

1. (2 points) Softmax maps logits to a vector of numbers that always sums to K (the number of classes).

   A. True

   B. False

   Solution. B

2. (2 points) For stochastic gradient descent to work well, the training data is often processed in a deterministic cyclic order.

   A. True

   B. False

   Solution. B

3. (2 points) With L_2 regularization $R(\theta) = \frac{1}{2}\|\theta\|_2^2$, the regularization part of the gradient descent update with learning rate η shrinks parameters by a factor $(1 - \eta\lambda)$ (ignoring the loss-gradient term).

   A. True

   B. False

   Solution. A

4. (2 points) In a convolutional layer, the same kernel weights are applied at all spatial locations of the input.

   A. True

   B. False

   Solution. A

5. (2 points) Batch normalization uses the same statistics during training and inference.

   A. True

   B. False

   Solution. B

6. (2 points) A recurrent neural network can, in principle, represent longer context than a fixed-window autoregressive model with the same hidden size.

   A. True

   B. False

   Solution. A

7. (2 points) A denoising autoencoder can reduce over-reliance on the identity map by training the model to reconstruct the clean input from a corrupted input x_{noisy}.

   A. True

B. False

Solution. A

8. (2 points) In an autoencoder, the output $\tilde{x}$ is trained to match a target label y rather than the input x.

   A. True

   B. False

   Solution. B

9. (2 points) If the VAE prior is $p(z) = \mathcal{N}(0, I)$, then after training we can generate samples by drawing $z \sim \mathcal{N}(0, I)$ and decoding.

   A. True

   B. False

   Solution. A

10. (2 points) In SimCLR, a decoder network is required during training in order to reconstruct the input.

    A. True

    B. False

    Solution. B

Part 2: Multiple Choice

Please clearly circle your answer. If you change your answer, please erase your previous marking cleanly. Each choice is graded separately and is worth 1 point.

11. (4 points) Which of the following statements about the Universal Approximation Theorem for two-layer neural networks is **true**? Assume the activation function ϕ_0 is continuous and non-polynomial. **Select all that apply.**

    A. Any continuous function defined on a bounded domain can be exactly represented by a two-layer neural network with finitely many hidden units.

    B. A multi-layer neural network defined on a bounded domain can be approximated arbitrarily well by a two-layer neural network with a finite number of hidden units.

    C. There exists an integer N such that a two-layer neural network with N hidden unit can approximate any continuous function arbitrarily well.

    D. The Universal Approximation Theorem only applies if the activation function is ReLU.

    Solution. B.

12. (4 points) Consider K-class classification with logits $z = f(\theta, x) \in \mathbb{R}^K$ and loss

$$-\ln \left[\mathrm{softmax}(z)\right]_y .$$

    Which statements are correct? **Select all that apply.**

    A. Minimizing this loss is equivalent to maximum likelihood estimation under the softmax model.

    B. This loss depends only on the predicted class $\arg\max_c z_c$ and not on the values of the other logits.

    C. This loss is commonly called cross-entropy loss.

    D. This loss can be used for multi-class logistic regression as well as neural networks.

    Solution. A,C,D.

13. (4 points) Which of the following is a typical disadvantage of full gradient descent compared with SGD or minibatch SGD on large datasets? **Select all that apply.**

    A. It requires a smaller learning rate than SGD to converge.

    B. Each update requires computing gradients over all n examples, which can be expensive when n is large.

    C. It may be impractical when the full dataset cannot be loaded or accessed in memory at once.

    D. It requires more memory to store model parameters.

    Solution. B, C.

14. (4 points) Which of the following statements is correct? **Select all that apply.**

    A. For a fixed minibatch size, doubling the learning rate always improves generalization.

    B. Increasing the minibatch size typically reduces the variance of the minibatch gradient estimate.

    C. Using a smaller learning rate can reduce oscillation but may slow down progress.

    D. A learning rate schedule that decreases η_t over time is often used to reduce the variance of updates later in training.

    Solution. B, C, D.

15. (4 points) Which of the following statements is correct about Adam versus AdamW? **Select all that apply.**

    A. Adam simply applies momentum to the gradients, without using any moving average of squared gradients.

    B. In AdamW, weight decay is applied as a separate shrinkage on parameters rather than being mixed into the adaptive gradient term.

    C. In Adam, the bias correction terms $\hat{m}_t$ and $\hat{v}_t$ are used because $m_0 = v_0 = 0$ makes early moving averages biased toward zero.

    D. When weight decay is nonzero, AdamW behaves the same as Adam because both apply regularization through the gradient update.

    Solution. B, C.

16. (4 points) Which of the following are typical advantages of convolutional layers over fully connected layers? **Select all that apply.**

    A. Fewer parameters due to parameter sharing.

    B. Ability to exploit local spatial structure.

    C. Exact invariance to all geometric transformations.

    D. Reduced memory usage for large images.

    Solution. A, B, D

17. (4 points) An input feature map has size $C_{\text{in}} \times 28 \times 28$. A convolution uses kernel size $k = 3$, stride $s = 1$, and **no padding** ($p = 0$), with $C_{\text{out}} = 16$. What is the output size?

    A. $16 \times 26 \times 26$

    B. $16 \times 28 \times 28$

    C. $16 \times 25 \times 25$

    D. $16 \times 14 \times 14$

    Solution. A.

18. (4 points) Which of the following correctly describe the effect of increasing stride in a convolution layer? **Select all that apply.**

    A. The spatial resolution of the output is not affected.

    B. The spatial resolution of the output is reduced.

    C. The number of output channels increases automatically.

    D. Some fine spatial details may be lost.

    Solution. B, D

19. (4 points) Which of the following statements about sequence models are correct? **Select all that apply.**

    A. In next-token prediction, each time step can be viewed as a supervised learning problem with the context as input and the next token as label.

    B. A bigram model is an example of a recurrent neural network.

    C. Tokenization changes the length of the input sequence seen by the model.

D. Increasing vocabulary size necessarily reduces training loss.

Solution. A, C

20. (4 points) Which of the following statements about tokenization are correct? **Select all that apply.**

A. Tokenization converts raw text into a sequence of integers.
B. Character-level tokenization always gives better performance than subword tokenization.
C. Subword tokenization helps handle rare or unseen words.
D. Changing the tokenizer changes the embedding and output layer sizes.

Solution. A, C, D

21. (4 points) Which of the following statements about sequence generation are correct? **Select all that apply.**

A. During generation, the model typically feeds its own previous predictions back as input.
B. Lower sampling temperature makes generated sequences more random.
C. Errors made early in generation can affect later outputs.
D. Greedy decoding (temperature 0) always maximizes the joint probability of the sequence.

Solution. A, C

22. (4 points) Which of the following statements is correct? **Select all that apply.**

A. In an encoder–decoder model, z needs to be a single fixed-length vector.
B. In image classification, a CNN is often used as an encoder to map an image x to an embedding z.
C. In image classification, the decoder should be a transposed-convolution network to output class logits.
D. In image classification, a small fully connected network can serve as a decoder that maps z to class logits.

Solution. B, D.

23. (4 points) Which of the following statements is correct? **Select all that apply.**

A. An autoencoder can be trained on unlabeled data by minimizing a reconstruction error $d(x, \tilde{x})$.
B. The main goal of an autoencoder is to output the correct class label for each input.
C. A low-dimensional z can act as a bottleneck that encourages compression of information about x.
D. If we increase the capacity of an autoencoder enough, it is guaranteed to discover a unique and interpretable representation z (independent of training details).

Solution. A, C.

24. (4 points) Which of the following statements is correct? **Select all that apply.**

A. In a denoising autoencoder, the target output is the noisy input x_{noisy} so that the model learns to preserve noise.
B. Adding noise during training makes reconstruction strictly easier, so training a denoising autoencoder should always be faster than training a standard autoencoder.
C. A denoising autoencoder is trained to map a corrupted input x_{noisy} back to a clean output close to x.
D. If the noise is Gaussian, then the decoder has to be linear for the method to work.

Solution. C.

25. (4 points) Which of the following statements is correct? **Select all that apply.**

   A. Transposed convolution is guaranteed to be the exact inverse operator of a standard convolution layer.

   B. With stride $s > 1$, standard convolution often reduces spatial resolution because outputs are computed on a sparser grid of positions.

   C. With stride $s > 1$, transposed convolution can increase spatial resolution.

   D. Max pooling preserves all information in the feature map because it keeps the maximum value in each window.

Solution. B, C.

26. (4 points) Which of the following statements is correct? **Select all that apply.**

   A. In a VAE, the encoder defines a distribution $q(z|x)$ rather than outputting only a point estimate for z.

   B. The KL regularization term encourages the decoder outputs to have zero mean and identity covariance in pixel space.

   C. In the VAE reparameterization $z = \mu(x) + \sigma(x) \odot \epsilon$, the noise ϵ is sampled independently of the input x.

   D. A standard auto-encoder always produces better samples than a VAE when sampling $z \sim \mathcal{N}(0, I)$.

Solution. A, C.

27. (4 points) Consider the PyTorch-style line `epsilon = torch.randn_like(std)` used in reparameterization. Which of the following statements is correct? **Select all that apply.**

   A. Sampling `epsilon` this way makes the KL term unnecessary, since the latent already follows $\mathcal{N}(0, I)$ by construction.

   B. Replacing `torch.randn_like(std)` by `torch.zeros_like(std)` keeps training unbiased for the ELBO objective.

   C. The sampled `epsilon` is treated as an external random input; gradients flow through `mu` and `std` in `mu + std*epsilon`.

   D. `torch.randn_like(std)` samples i.i.d. standard normal noise with the same shape as `std`.

Solution. C, D.

28. (4 points) Consider a conditional VAE modeling $p(y|x)$ with encoder $q(z|x, y)$ and decoder $p(y|x, z)$. Which of the following statements is correct? **Select all that apply.**

   A. Conditioning x can be fed to both encoder and decoder to model $q(z|x, y)$ and $p(y|x, z)$.

   B. In a conditional VAE, z can capture information about y that is not determined by x.

   C. The correct variational distribution for a conditional VAE is $q(z|x)$, since y should not be used during training.

   D. After training, to generate y given x, we can sample $z \sim \mathcal{N}(0, I)$ and decode using $\mathrm{dec}(x, z)$.

Solution. A, B, D.

29. (4 points) Which statements correctly describe the batch construction in SimCLR? **Select all that apply.**

A. From a minibatch of N original examples, we create $2N$ augmented views by sampling two augmentations per original.

B. The positive pairs are chosen as the two most similar views under the current encoder, since we do not have labels.

C. Each augmented view has exactly one positive in the batch: the other view made from the same original example.

D. For a fixed anchor view i, all other $2N - 1$ views (except itself) appear in the denominator of $\ell(i, j)$.

Solution. A, C, D.

30. (4 points) Which statements about the role of augmentations in SimCLR are correct? **Select all that apply.**

A. Strong augmentations reduce downstream performance because they make the optimization problem too noisy.

B. If two augmented views are treated as a positive pair, the model is trained to make their embeddings close.

C. Strong augmentations can reduce reliance on low-level pixel cues by forcing invariance to larger appearance changes that keep semantics.

D. If we never use color distortion or crop, then the model is still forced to be invariant to those transformations.

Solution. B, C.

This page is intentionally left blank to accommodate work that wouldn't fit elsewhere and/or scratch work.