24-789 Intermediate Deep Learning
Spring 2025
Midterm
4/14 12:00 AM - 4/15 11:59 PM (EDT)

Name (Print): ______________________________

Andrew ID: ______________________________

This exam contains 7 pages (including this cover page) and 3 problems. Check to see if any pages are missing. Enter all requested information on the top of this page, and put your name on the top of every page, in case the pages become separated.

You are required to show your work on each problem on this exam. The following rules apply:

- **Organize your work**, in a reasonably neat and coherent way, in the space provided. Work scattered all over the page without a clear ordering will receive very little credit.
- **Mysterious or unsupported answers will not receive full credit**. A correct answer, unsupported by calculations, explanation, or algebraic work will receive no credit; an incorrect answer supported by substantially correct calculations and explanations might still receive partial credit.
- If you need more space, use the back of the pages; clearly indicate when you have done this.

Do not write in the table to the right.

Problem	Points	Score
1	40	
2	30	
3	30	
Total:	100	

1. (40 points) Short Answer Questions

   (a) (4 points) How do the underlying mechanisms of diffusion models differ from those of GANs and VAEs? What are the respective advantages and limitations of each approach in terms of stability and efficiency of training, as well as efficiency and output quality of data generation?

   (b) (4 points) What is the difference between an autoencoder and a variational autoencoder (VAE)? How does the latent space differ between the two types of autoencoders?

   (c) (4 points) What are the main limitations of RNNs and LSTMs that the Transformer architecture mitigates?

   (d) (4 points) What's the main difference between Transformer's encoder and decoder?

   (e) (4 points) Explain the role of Generator and Discriminator in Generative Adversarial Networks (GANs). How do GANs generate data at inference time? How does this differ from the approach used by Variational Autoencoders (VAE) to generate new data?

   (f) (4 points) How does a Bidirectional GAN differ from a vanilla GAN? How does the structure map to the latent space?

   (g) (4 points) What is the point of including the KL divergence loss in the objective function while training a VAE? What happens if we assign too much or too little weight on this loss term?

   (h) (4 points) Explain what happens if you over-train the discriminator in a GAN? What is a way to prevent over-training of the discriminator?

   (i) (4 points) Mode collapse is a possible problem when generating data using GANs. Explain what that is, when it might happen, and how one can fix it.

   (j) (4 points) What is the role of transposed convolutional layers in a Deep Convolutional Generative Adversarial Network (DCGAN)? How do they enable the generator network to produce high-resolution images?

2. (30 points) Calculation

Please give your calculation steps and necessary descriptions. You will lose points if only results are given.

(a) (10 points) In this question you are asked to calculate a causal self-attention by hand. The input x_i is the encoding of each word in the sentence, which contains three words:

$$X = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0.2 & 0.8 \\ 0.5 & 0.5 \\ 1 & 0.5 \end{bmatrix}. \tag{1}$$

The following weight matrices are applied to the input features to project them into query, key, and value:

$$Q = XW_q; \quad K = XW_k; \quad V = XW_v \tag{2}$$

$$\text{where: } W_q = \begin{bmatrix} 0.5 & 0.5 \\ 0 & 1 \end{bmatrix}, \quad W_k = \begin{bmatrix} 1 & 0 \\ -0.5 & 0.5 \end{bmatrix}, \quad W_v = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{3}$$

The attention used here will be standard scaled-dot product attention (with scaling factor d_k set to 1), and is causal attention (as in the Transformer decoder), which means that token cannot attend with the token that appears after it, e.g. x_0 cannot attend with x_2 but x_2 can attend with x_0. The attention here will contain only one head.

Please calculate the final output features of this attention layer, $O = A(Q, K)V$, where A is the causal self-attention (O should also be a 3×2 matrix).

(b) (10 points) In this question you are asked to calculate the FID scores for two sets of fake images with respect to a set of real images on dummy values with 4 features to find out which set of generated images would be considered better.

The FID score is a measure of the similarity between the feature distributions of the real and generated image sets and can be used as a way to evaluate the performance of generative models like GANs. In reality, to calculate the FID score, you need to use an Inception network to extract 1000s of features from both the real images in your dataset and the generated images produced by your GAN.

For this dummy example, The features have already been extracted for you and the specifications are as follows:

Real Image set:

$$\mu_{real} = \begin{bmatrix} 0.5 & 0.2 & 0.7 & 0.4 \end{bmatrix} \Sigma_{real} = \begin{bmatrix} 0.2 & 0.1 & 0.05 & 0.02 \\ 0.1 & 0.3 & 0.1 & 0.05 \\ 0.05 & 0.1 & 0.4 & 0.1 \\ 0.02 & 0.05 & 0.1 & 0.2 \end{bmatrix} \tag{4}$$

Fake Images Set I:

$$\mu_1 = \begin{bmatrix} 0.4 & 0.3 & 0.5 & 0.6 \end{bmatrix} \Sigma_1 = \begin{bmatrix} 0.7 & 0.1 & 0.05 & 0.02 \\ 0.1 & 0.2 & 0.1 & 0.05 \\ 0.05 & 0.1 & 0.3 & 0.1 \\ 0.02 & 0.05 & 0.1 & 0.15 \end{bmatrix} \tag{5}$$

Fake Images Set II:

$$\mu_2 = \begin{bmatrix} 0.7 & 0.1 & 0.3 & 0.5 \end{bmatrix} \Sigma_2 = \begin{bmatrix} 0.4 & 0.05 & 0.05 & 0.01 \\ 0.05 & 0.3 & 0.1 & 0.05 \\ 0.05 & 0.1 & 0.2 & 0.05 \\ 0.01 & 0.05 & 0.05 & 0.2 \end{bmatrix} \tag{6}$$

Please report your FID1 and FID2 scores along with calculations. Also mention which set of generated images is better in your opinion based on FID score.

Note: μ here refers to the mean vector and Σ here refers to the covariance matrix of the feature representations of a set of images. μ contains feature-wise means, where each element represents the mean value of a particular feature across all images in the set. The covariance matrix Σ is a square matrix that describes the relationships between different features in the set.

(c) (10 points) In this question, you are given the image x_0 and schedule β. Perform two iterations of the forward diffusion process and report the noisy image output x_1 and x_2.

The forward diffusion equation is given below:

$$q\left(x_t \mid x_{t-1}\right) = N\left(x_t; \sqrt{1-\beta_t} x_{t-1}, \beta_t I\right) \tag{7}$$

Image:

$$x_0 = \begin{bmatrix} 0.8 & 0.5 & 0.9 \\ 0.1 & 1 & 0.6 \\ 0.4 & 0.6 & 0.3 \end{bmatrix} \tag{8}$$

Standard deviation β_0: 0.0001 and β_{1000}: 0.02 Use a linear schedule to estimate the values of β_1 and β_2. The noise at t=1 and t=2 are given as:

$$\epsilon_1 = \begin{bmatrix} 1.0 & -1.0 & 1.0 \\ -1.0 & 1.0 & 1.0 \\ 1.0 & -1.0 & 1.0 \end{bmatrix} \quad \epsilon_2 = \begin{bmatrix} 2.0 & -2.0 & 2.0 \\ -2.0 & 2.0 & 2.0 \\ 2.0 & -2.0 & 2.0 \end{bmatrix}$$

Note: Use the reparameterization trick

3. (30 points) Multiple choice questions

Please select the ONE correct answer! You don't need to give the explanation of your choices.

Question 1: Which of the following is true about Diffusion Models.

A) It is theoretically impossible to construct a diffusion model that does not use conjugate prior and posterior distributions.

B) Conjugate prior and posterior distributions allow us to evaluate the loss function rather than rely on sampling techniques.

C) Latent Diffusion Models must learn the encoding and diffusion process simultaneously due to numerical instability during reverse diffusion sampling.

D) In Cascading Diffusion Models, conditioning augmentation refers to augmenting the initial step at each level of the cascade.

Question 2: Suppose you are training a VAE model. Even though the loss function converges after a good amount of iterations, you observe that the reconstructed samples are very similar even for different input samples. What is a possible way of addressing this issue?

A) Decrease the weight of the KL divergence term in the loss function and retrain the model from scratch.

B) Reduce the dimensionality of the input data and retrain the model from scratch.

C) Increase the learning rate and continue training the current model.

D) None of the Above

Question 3: Which of the following is **NOT** a benefit of using attention mechanisms in Transformer models?

A) Improved performance on tasks that require modeling long-range dependencies

B) To have a scalable model in which performance improved when having more parameters

C) Ability to process inputs of varying lengths

D) To reduce number of model parameters

Question 4: Which of the following statements about self-attention in the Transformer architecture is **TRUE**?

A) Self-attention is only used to process the input sequence and is not used during the generation of new outputs.

B) Self-attention computes a weighted sum of the values using the dot product (normalized with softmax) of the queries and keys as the weights.

C) Self-attention is only used in the encoder layers of the Transformer, not in the decoder layers.

D) Self-attention cannot be used with variable-length input sequences.

Question 5: Which of the following statements about the attention mechanism in the Transformer architecture is **FALSE**?

A) The attention mechanism is a form of dynamic weight learning that allows the model to learn to selectively attend to different parts of the input sequence when computing each output.

B) The dot product attention used in the Transformer has a quadratic complexity with respect to the number of input tokens.

C) The multi-head attention mechanism in the Transformer allows the model to attend to different sequences within the same batch.

D) The attention mechanism can be modified to allow the model to attend to different modalities, such as image features.

Question 6: Which of the following is **NOT** a commonly used Loss Function in Generative Adversarial Networks (GANs)?

A) Binary cross-entropy.

B) Huber Loss.

C) Kullback-Leibler Divergence.

D) Wasserstein Distance.

Question 7: What is the role of noise in generative networks?

A) To decrease the loss in the minimax game.

B) To ensure KL divergence is large enough.

C) To ensure that our training is random.

D) Enable the creation of new and unique outputs.

Question 8: What is the role of a kernel in the discriminator of a GAN?

A) Down sample the images from a feature map.

B) Up sample the images from a feature map.

C) Generate new information from the latent space.

D) Reduce the size of an image without any learned parameters.

Question 9: Why do we use a variance schedule in the forward process of diffusion models?

A) To control the amount of noise added at each diffusion step.

B) To adjust the learning rate of the optimization process.

C) To increase the diversity of the generated image.

D) To regulate the color balance and saturation of the image.

Question 10: In the context of attention mechanisms in neural networks, what is the purpose of the query, key, and value vectors?

A) Query vectors encode the information from the input sequence, key vectors determine the relevance of each element, and value vectors store the learned context.

B) Query vectors determine the relevance of each element, key vectors encode the information from the input sequence, and value vectors store the learned context.

C) Query vectors store the learned context, key vectors determine the relevance of each element, and value vectors encode the information from the input sequence.

D) Adding attention to the zt and ft activations of a GRU to add attention information alongside long and short term memory.