CSE 493 G1/ 599 G1
Deep Learning
Winter 2024 Quiz 5

March 7, 2024

Full Name: _____

UW Net ID: _____

Question		Score
True/False	(4 pts)	
Multiple Choice	(8 pts)	
Short Answer	(8 pts)	
Total	(20 pts)	

Welcome to the CSE 493 G1 Quiz 5!

- The exam is 20 min and is **double-sided**.
- No electronic devices are allowed.

I understand and agree to uphold the University of Washington Honor Code during this exam.

Signature: _____ Date: _____

Good luck!

This page is left blank for scratch work only. DO NOT write your answers here.

1 True / False (4 points) - Recommended 4 Minutes

Fill in the circle next to True or False, or fill in neither. Fill it in completely like this: ●. No explanations are required.

Scoring: Correct answer is worth 1 points.

1.1 GPT is an auto-regressive model, but BERT is not.

◯ True

◯ False

1.2 Both VAEs and autoregressive approaches rely on explicit density functions, but diffusion models use implicit density functions

◯ True

◯ False

1.3 Flamingo is an encoder-decoder model.

◯ True

◯ False

1.4 VAEs may be a better choice than diffusion models if you wish to generate images conditioned on some text, as diffusion models are unable to do this.

◯ True

◯ False

2 Multiple Choices (8 points) - Recommended 8 Minutes

Fill in the circle next to the letter(s) of your choice (like this: ●). No explanations are required. **Choose ALL options that apply.**

Each question is worth 2 points and the answer may contain one or more options. Selecting all of the correct options and none of the incorrect options will get full credits. For questions with multiple correct options, each incorrect or missing selection gets a 1-point deduction (up to 2 points).

2.1 You are using a decoder-only language model to generate a 3 word sentence (i.e. "I love pizza"). At each step, you input the portion of the sequence that has already been generated and then output a probability distribution (using softmax) over the potential next words. You then sample from this distribution and add this word to the end of your sequence. You repeat this until you have your three words. Your first generated word is "I". The conditional probabilities calculated by the language model over some potential next words given previous sequences are listed below. $P(\text{"love snacks"}|\text{"I"})$ means the probability that the sequence "I love snacks" will be generated, given that "I" is already generated.

$$P(\text{"love"}|\text{"I"}) = 0.5 \qquad P(\text{"ate"}|\text{"I"}) = 0.2 \qquad P(\text{"hate"}|\text{"I"}) = 0.1$$
$$P(\text{"pizza"}|\text{"I love"}) = 0.5 \qquad P(\text{"pizza"}|\text{"I ate"}) = 0.1$$
$$P(\text{"snacks"}|\text{"I love"}) = 0.4 \qquad P(\text{"snacks"}|\text{"I ate"}) = 0.8$$

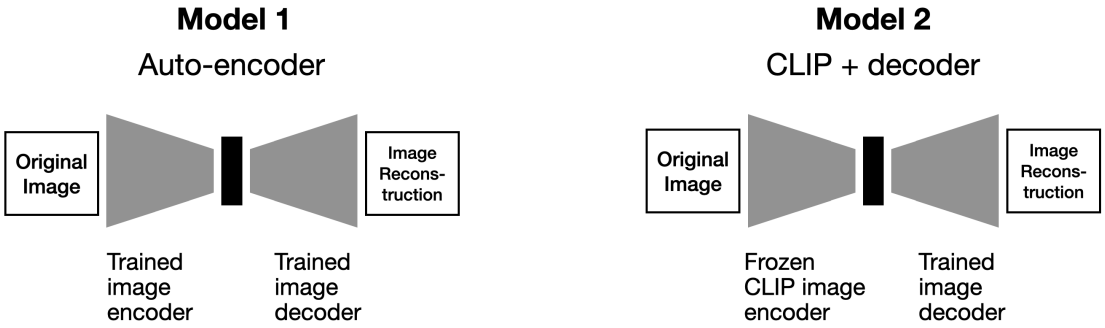
- ◯ A: $P(\text{"love pizza"}|\text{"I"}) = 0.25$
- ◯ B: If you greedily decode (or pick the most likely word at each step), you will end up with the sequence 'I love snacks'
- ◯ C: $P(\text{"pizza"}|\text{"I love"}) + P(\text{"pizza"}|\text{"I ate"}) + P(\text{"pizza"}|\text{"I hate"})$ must be less than 1.
- ◯ D: $P(\text{"saw"}|\text{"I"}) < 0.2$
- ◯ E: $log(P(\text{"I ate snacks"})) < log(P(\text{"I love snacks"}))$

2.2 Answer the following questions about prompting

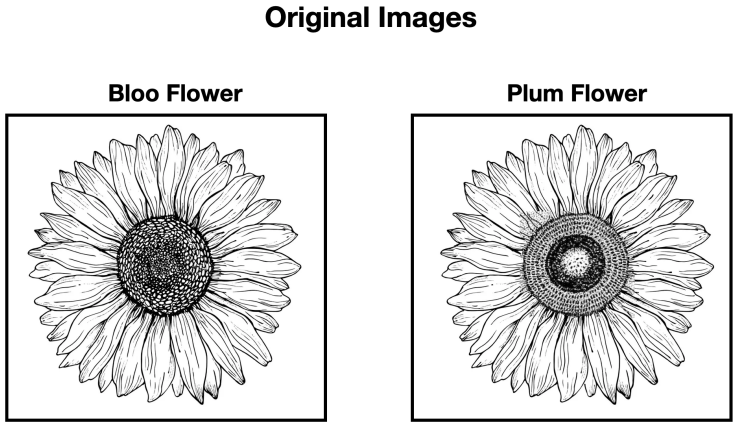
- ◯ A: Zero-shot refers to when a language model which is able to complete a task without any description of what that task is.
- ◯ B: Chain-of-thought prompting is a type of in-context learning
- ◯ C: The GPT-3 paper found that more in-context example could help improve accuracy for many tasks
- ◯ D: The authors of CLIP found that they were able to improve ImageNet accuracy by adding "A photo of a ___"before the name of the object category (i.e. "A photo of a dog" instead of just "dog"), but only if they first fine-tuned on captions that contained the phrase "A photo of a ___"

3 Short Answers (8 points) - Recommended 8 Minutes

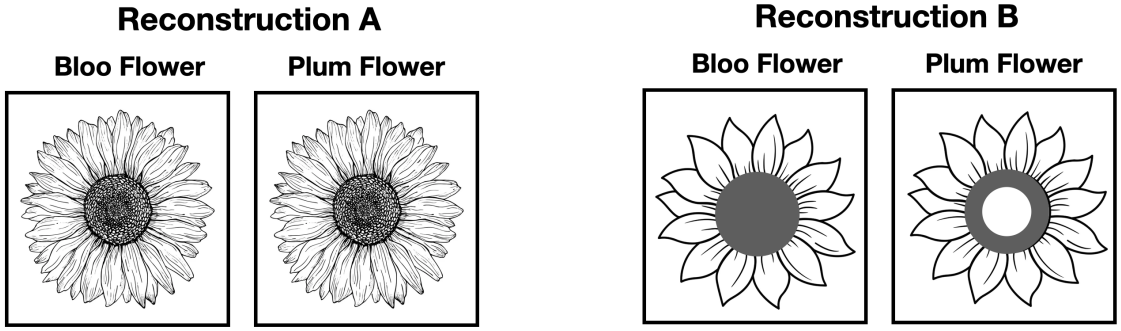
Please make sure to write your answer only in the provided space.
Consider the following 2 models. Model 1 is an autoencoder, where both the encoder and decoder have been trained with L2 reconstruction loss. Model 2 is a modified autoencoder where the encoder portion is the image encoder from CLIP that has been frozen (so the weights have not been modified after it was trained using the CLIP training objective). The decoder is then trained using the L2 reconstruction loss to make the best reconstruction possible given the embeddings generated by the CLIP encoder.



After training, you send the following two images through each model. One image is of a Bloo flower and one image is of a Plum flower. These flowers can have many different appearances, but Bloo flowers always have a dark center and Plum flowers always have a light center. ***This is the only way to tell them apart.***



One of the models outputs Reconstruction A and one of the models output reconstruction B.



3.1 Match Reconstructions (1 point)

Which of the following is true?

◯ A: Reconstruction A is from Model 1 and Reconstruction B is from Model 2

◯ B: Reconstruction B is from Model 1 and Reconstruction A is from Model 2

3.2 Justify (2 points)

Explain your reasoning for your above answer.

3.3 Accuracy (1 point)

Which of the following is true?

◯ A: Embeddings from Model 1 are better for Bloo/Plum Flower Classification

◯ B: Embeddings from Model 2 are better for Bloo/Plum Flower Classification

3.4 Justify (2 points)

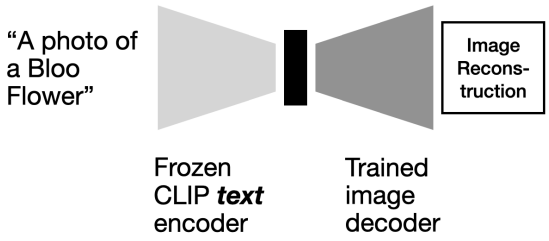
Explain your reasoning for your above answer.

3.5 Patching together models (2 points)

You then create a third model by taking the CLIP text encoder (not image encoder), generating a text embedding and then passing this into the image decoder from Model 2.

Model 3

CLIP ***text*** encoder + Model 2 ***image*** decoder



You do not fine-tune/train this model at all, but you find when you pass the sentence "A photo of a Bloo Flower" into the text encoder, the image decoder produces a rough image of a Bloo Flower. Why does this work at all with no additional training?