CSE 493 G1/ 599 G1
Deep Learning
Spring 2024 Quiz 4

SOLUTIONS

Mar 1, 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 4!

- 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!

Mean: 14.73
Median: 14.5
Stdev: 4.05

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. Incorrect answer and leaving it blank will get 0 points.

1.1 In transfer learning with CNNs, when the target data distribution is similar to the original pretraining data distribution, it's common to replace and retrain the last fully connected layer while keeping earlier layers frozen.

○ True

○ False

SOLUTION:
True

1.2 Masked language modeling is a type of self-supervised learning in which the model learns to predict masked text without explicit labels or annotations.

○ True

○ False

SOLUTION:
True

1.3 In self-supervised learning, excellent performance on a pretext task always guarantees similarly high performance on downstream tasks.

○ True

○ False

SOLUTION:
False.

1.4 Transposed convolution is often used for reducing the spatial size of the feature map.

○ True

○ False

SOLUTION:
False.

2 Multiple Choices (8 points) - Recommended 6 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 4 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 2-point deduction (up to 4 points).

2.1 Select all statements that are true about LLMs.

○ A: LLMs are pre-trained on large datasets and can perform a wide range of language tasks without task-specific training data.

○ B: BERT is primarily trained with masked language modeling and next sentence prediction .

○ C: ELMo uses a bi-directional model for next-word prediction, which helps in capturing the context from both previous and following words

○ D: Chain of thought prompting is a technique where LLMs are fine-tuned to generate intermediate reasoning steps, enhancing their problem-solving abilities.

○ E: None of the above.

SOLUTION:
A, B, C. (D) Chain of thought prompting does not require additional fine-tuning.

2.2 Select all statements that are true about CLIP model.

○ A: CLIP uses a contrastive loss function to align text and image embeddings.

○ B: CLIP requires image labels for training.

○ C: CLIP is trained to maximize the cosine similarity between the embeddings of matching text-image pairs.

○ D: CLIP model is trained with image transformation pretext tasks such as rotation prediction and inpainting to learn effective feature representations.

○ E: To use CLIP models for image classification, it is necessary to fine-tune them on the downstream datasets.

SOLUTION:
A, C. (B) CLIP does not require image labels, (D) CLIP model is trained with contrastive loss, (E) CLIP can perform zero-shot image classification without further training

Grading Note: The following clarification was issued on the discussion board in response to student questions regarding why B was not considered correct:

> In the context of self-supervised learning and CLIP models, image labels refer to traditional categorical image classes. Text descriptions do NOT count as labels.

3 Short Answers (8 points) - Recommended 10 Minutes

Please make sure to write your answer only in the provided space.

The Contrastive Language-Image Pre-training (CLIP) model uses a large-scale dataset of images and corresponding textual descriptions to learn visual concepts from natural language supervision. It's trained to predict which images are paired with which texts, effectively learning to understand both modalities. Denote $\mathcal{X}$ as the image space and $\mathcal{Y}$ as the text space. You are given a dataset consisting of images-text pairs $\{x_i \in \mathcal{X}, y_i \in \mathcal{Y}\}_{i=1}^n$. Your task is to train image encoder $f_\theta : \mathcal{X} \to \mathcal{E}$ and text decoder $g_\psi : \mathcal{Y} \to \mathcal{E}$ such that they can align image and text representations in a shared embedding space $\mathcal{E}$.

Given a mini-batch of image-text pairs $(x_1, y_1), (x_2, y_2), \cdots, (x_B, y_B)$, we write contrastive loss function for each image x_i with respect to all texts in the batch as follows:

$$\mathcal{L}_{\text{image }}(x_i) := -\log \frac{\exp\left(\cos\left(f_\theta\left(x_i\right), g_\psi\left(y_i\right)\right)\right)}{\sum_{k=1}^B \exp\left(\cos\left(f_\theta\left(x_i\right), g_\psi\left(y_k\right)\right)\right)} .$$

where $\cos(u, v) := \frac{u^\top v}{\|u\|\|v\|}$.

Similarly, the contrastive loss for each text y_i with respect to all images in the batch is:

$$\mathcal{L}_{\text{text }}(y_i) := -\log \frac{\exp\left(\cos\left(g_\psi\left(y_i\right), f_\theta\left(x_i\right)\right)\right)}{\sum_{k=1}^B \exp\left(\cos\left(g_\psi\left(y_i\right), f_\theta\left(x_k\right)\right)\right)} .$$

The total loss for the mini-batch is the average of the losses for all images and texts:

$$\mathcal{L}(\theta, \psi) := \frac{1}{2} \sum_{i=1}^B \left[\mathcal{L}_{\text{image }}(x_i) + \mathcal{L}_{\text{text }}(y_i)\right] .$$

1. (4 points) Suppose you have perfect image encoder and text encoder that align the matching image and text pair, i.e. $\cos(f_\theta(x_i), g_\psi(y_i)) = 1$ and $\cos(f_\theta(x_i), g_\psi(y_k)) = -1$ for $k \neq i$, does the contrastive loss $\mathcal{L}_{\text{image}}(x_i)$ go to zero? Briefly justify your answer.

   SOLUTION:
   No. Under perfect image and text encoder assumption, we have

   $$\mathcal{L}_{\text{image }}(x_i) = -\log \frac{e}{e + \frac{B-1}{e}}$$

   For $B > 2$, $\mathcal{L}_{\text{image }} > 0$. This means the total contrastive loss is lower bounded by some positive number that increases with batch size B.

2. (4 points) Given pre-trained image encoder $f_\theta(\cdot)$ and text encoder $g_\psi(\cdot)$, how would you use them to build a zero-shot image classifier? Assume we have object categories $c_0, c_1, \cdots, c_M$, given a test image x, write an $\arg\max_c$ function that will return the correct class using $f_\theta(\cdot)$ and $g_\psi(\cdot)$.

   SOLUTION:
   For test image x, the model generates an image embedding $f_\theta(x)$. For each category c_j, a textual description $\tilde{c}_j$ (e.g., ”a photo of a cat”) is encoded into a text embedding $g_\psi(\tilde{c}_j)$. The category whose text embedding is closest to the image embedding (as measured by cosine similarity) is the predicted category for the image, e.g. $\hat{c}(x) = \arg\max_c \cos(f_\theta(x), g_\psi(\tilde{c}))$

3. [Extra Credit 4 points] Please rewrite $\mathcal{L}_{\text{image}}\ (x_i)$ to show that it is equivalent to smooth (i.e. differentiable) version of Triplet loss:

$$\mathcal{L}_{\text{image}}\ (x_i) \approx \max\left(0, \max_{j\neq i}\{\cos(f_\theta(x_i), g_\psi(y_j))\} - \cos(f_\theta(x_i), g_\psi(y_i))\right)\ .$$

   You may use the following continuous relaxation in your proof:

$$\begin{aligned}\max(x, y) &= \lim_{k\to\infty}\frac{1}{k}\ln(e^{kx} + e^{ky}) \approx \log(e^x + e^y)\\ \max(x_1, \ldots, x_n) &= \lim_{k\to\infty}\frac{1}{k}\ln(\sum_{i=1}^{n} e^{kx_i}) \approx \log(\sum_{i=1}^{n} e^{x_i})\ .\end{aligned}$$

   SOLUTION:

$$\begin{aligned}\mathcal{L}_{\text{image}}\ (x_i) &= -\log\frac{\exp\left(\cos\left(f_\theta\left(x_i\right), g_\psi\left(y_i\right)\right)\right)}{\sum_{k=1}^{B}\exp\left(\cos\left(f_\theta\left(x_i\right), g_\psi\left(y_k\right)\right)\right)}\\ &= \log\left(1 + \frac{\sum_{k\neq j,k=1}^{B}\exp\left(\cos\left(f_\theta\left(x_i\right), g_\psi\left(y_k\right)\right)\right)}{\exp\left(\cos\left(f_\theta\left(x_i\right), g_\psi\left(y_i\right)\right)\right)}\right)\\ &\approx \max\left(0, \log\frac{\sum_{k\neq j,k=1}^{B}\exp\left(\cos\left(f_\theta\left(x_i\right), g_\psi\left(y_k\right)\right)\right)}{\exp\left(\cos\left(f_\theta\left(x_i\right), g_\psi\left(y_i\right)\right)\right)}\right)\\ &= \max\left(0, \log\left(\sum_{k\neq j,k=1}^{B}\exp\left(\cos\left(f_\theta\left(x_i\right), g_\psi\left(y_k\right)\right)\right)\right) - \cos\left(f_\theta\left(x_i\right), g_\psi\left(y_i\right)\right)\right)\\ &\approx \max\left(0, \max_{k\neq j}\cos\left(f_\theta\left(x_i\right), g_\psi\left(y_k\right)\right) - \cos\left(f_\theta\left(x_i\right), g_\psi\left(y_i\right)\right)\right)\ .\end{aligned}$$