**Upload your project essay and related figures in MyCourses by 11:59PM, 11/01/2024**

This project is not a teamwork project. You complete this project by yourself. Grading will be strictly followed by the grading rubrics in this document.

**Project 2. Multimodal AI – OpenAI CLIP**

The project contains 2 sub-tasks that are related to multimodal AI with OpenAI CLIP.

**Task 1:** Study the paper attached to Project 2. The CLIP paper is “Learning Transferable Visual Models From Natural Language Supervision”.

Read this paper and understand some questions like what the CLIP (Contrastive Language-Image Pre-training) is.

After you read the paper in task 1, you would write a short essay. Write your essay in MS Word. Use the “Times New Roman” font and 12 font size. Up to 3 pages. Grammar and document format will be considered in grading.

The structure of the essay should follow the rules below.

1. Introduction

Answer questions:

* 1. What are the pre-training methods?
  2. What are the contributions of this CLIP paper?

1. Overview

Briefly state the history of natural language supervision, creating a sufficiently large dataset, selecting an efficient pre-training method, and choosing and scaling a model.

1. Workflow of the CLIP in Figure 1

(3.1). What is a contrastive pre-training? Use 3 to 5 sentences to describe it.

(3.2). How to create a dataset classifier from label text? Use 3 to 5 sentences to describe it.

(3.3). How to do zero-shot prediction in CLIP? Use 3 to 5 sentences to describe it.

1. Zero-Shot Transfer

(4.1). What is the zero-shot transfer? Use 3 to 5 sentences to describe it.

(4.2). How to do the zero-shot transfer by using the CLIP model? Use 3 to 5 sentences to describe it.

(4.3). In comparison to the traditional method such as Visual N-Grams, what are the advantages of the CLIP for zero-shot transfer? Use 3 to 5 sentences to describe it.

(4.4). Summarize the performance of the zero-shot CLIP. Use 3 to 5 sentences to describe it.

1. Distribution Shift

(5.1). What is the distribution shift problem of a model? Use 3 to 5 sentences to describe it.

(5.2). Why the CLIP model is robust to natural distribution shift? Use 3 to 5 sentences to describe it.

1. Comparison to Human Performance

(6.1). List all comparisons between the CLIP model and humans in this paper. For each comparison, please use 2 – 3 sentences to describe it.

1. Limitations and Broader Impacts

(7.1). List all limitations of the CLIP model presented in this paper. For each limitation, please use 2 – 3 sentences to describe it.

(7.2). List all broader impacts provided by the CLIP model. For each broader impact, please use 2 – 3 sentences to describe it.

1. Limitations and Future Work

(7.1). List all limitations of the CLIP model presented in this paper. For each limitation, please use 2 – 3 sentences to describe it.

(7.2). List all future works presented by authors. For each future work, please use 2 – 3 sentences to describe it.

**Task 2:** Using Python to run the CLIP model and experience multimodal AI using your own datasets.

(1). Download the Notebook file “Interacting with CLIP.ipynb” from here: <https://colab.research.google.com/github/openai/clip/blob/master/notebooks/Interacting_with_CLIP.ipynb>

You can also find this Notebook file attached to this project. The repo of the pre-trained OpenAI CLIP model can be found here:

<https://github.com/OpenAI/CLIP>

(2). Open the Notebook file in Google Colab. Colab is free cloud computing platform. You do not need to pay for it. If you do not have a Google Account for using Colab freely, please create a Google account for free.

If you never used Colab before, please study the basic usage of Colab by reading the tutorial: <https://www.tutorialspoint.com/google_colab/index.htm>

(3). Run the Notebook file “Interacting with CLIP.ipynb” on Colab. Experience the process of the CLIP model.

(4). Use your own dataset for evaluating and testing the CLIP model. Go to <https://cocodataset.org/#explore>, and download 10 images for building your image database. For example, if input the “table” for exploration, many images are displayed. Select 10 images and download them to your own computer. You can choose any one keyword such as “cat”, “car”, “bowl”, “apple”, “clock”, and other words included in the COCO database. Most popular concepts are included in the following figure. After you select 10 images, put those images into one folder named “image10”. Note that the selected images must be different from the images used in the original code of the original “Interacting with CLIP.ipynb”. You can name your code file “New Interacting with CLIP.ipynb”.

A group of icons with black and white squares

Description automatically generated

(5). Use these 10 images you selected from the COCO dataset and write a prompt for each image. Then, input the pairs of images and prompts into the Colab code of CLIP model in step (3) above. Replace the images and prompts in the original code of “Interacting with CLIP.ipynb” by the COCO images and prompts you generated. Then run the code and create the new matrix of the Cosine similarity between text and image feature. Your matrix structure is similar to the matrix below, but your matrix will use new images from COCO dataset and your own prompts. Note that you have 10 images acquired from COCO dataset, the new matrix you generate should be 10 by 10 size, rather than 8 by 8 size in the sample figure below.

A screenshot of a computer screen

Description automatically generated

(6). Finally, create the probability diagrams for each image of 10 COCO images. Your probability diagram may be similar to the figure below.

A screenshot of a computer

Description automatically generated

(6). Upload the matrix figure of the step (5) and the probability diagram of the step (6) with the essay written to MyCourses. For security of API Key information, do not upload the code file “New Interacting with CLIP.ipynb”.

**Upload the following documents (3 items in total) on MyCourses by 11:59PM, 11/01/2024 for grading evaluation. This project is not a teamwork project. You complete this project by yourself.**

1. **Task 1: one essay which is consistent with the requirements as presented in this project assignment instruction**. Grammar and document format will be considered in grading.
2. **Task 2: two figures of Task 2: new matrix of Cosine similarity between texts and image features of step (5) of Task 2, and probability diagram of step (6) of Taks 2. Note that DO NOT submit your code file “New Interacting with CLIP.ipynb” because it contains your API Key information. Keep API Key information security by yourself.**

**Grading Rubrics**

**Task 1: Essay (60 Points)**

**1. Introduction (10 Points)**

* **Pre-Training Methods (5 Points)**
  + Full marks for clear, accurate, and concise explanation of pre-training methods.
* **Contributions of the CLIP Paper (5 Points)**
  + Full marks for identifying key contributions of the CLIP paper and explaining their significance.

**2. Overview (10 Points)**

* **History of Natural Language Supervision (2 Points)**
  + Full marks for a brief, accurate overview.
* **Creating a Sufficiently Large Dataset (2 Points)**
  + Full marks for clear explanation.
* **Selecting an Efficient Pre-Training Method (2 Points)**
  + Full marks for clear explanation.
* **Choosing and Scaling a Model (2 Points)**
  + Full marks for clear explanation.

**3. Workflow of the CLIP in Figure 1 (10 Points)**

* **Contrastive Pre-Training (3 Points)**
  + Full marks for a clear, concise description.
* **Creating a Dataset Classifier from Label Text (3 Points)**
  + Full marks for clear, concise description.
* **Zero-Shot Prediction (4 Points)**
  + Full marks for a clear, concise description.

**4. Zero-Shot Transfer (10 Points)**

* **Zero-Shot Transfer Definition (2 Points)**
  + Full marks for a clear, concise definition.
* **Zero-Shot Transfer with CLIP (2 Points)**
  + Full marks for a clear, concise explanation.
* **Advantages of CLIP for Zero-Shot Transfer (3 Points)**
  + Full marks for a clear, concise comparison.
* **Performance Summary of Zero-Shot CLIP (3 Points)**
  + Full marks for a clear, concise summary.

**5. Distribution Shift (8 Points)**

* **Distribution Shift Problem (4 Points)**
  + Full marks for a clear, concise explanation.
* **Robustness of CLIP to Natural Distribution Shift (4 Points)**
  + Full marks for a clear, concise explanation.

**6. Comparison to Human Performance (4 Points)**

* **Comparisons between CLIP and Humans (4 Points)**
  + Full marks for listing and accurately describing comparisons.

**7. Limitations and Broader Impacts (8 Points)**

* **Limitations (4 Points)**
  + Full marks for listing and accurately describing limitations.
* **Broader Impacts (4 Points)**
  + Full marks for listing and accurately describing broader impacts.

**8. Limitations and Future Work (5 Points)**

* **Limitations (2 Points)**
  + Full marks for listing and accurately describing limitations.
* **Future Works (3 Points)**
  + Full marks for listing and accurately describing future works.

**9. Grammar and Document Format (5 Points)**

* **Grammar (2.5 Points)**
  + Full marks for correct grammar and syntax.
* **Document Format (2.5 Points)**
  + Full marks for following the specified format (Times New Roman, 12-point font, up to 3 pages).

**Task 2: Practical Implementation (40 Points)**

**1. Running the CLIP Notebook on Colab (10 Points)**

* **Successful Execution (5 Points)**
  + Full marks for successfully running the provided notebook.
* **Understanding of Process (5 Points)**
  + Full marks for demonstrating a clear understanding of the steps involved.

**2. Using Own Dataset (10 Points)**

* **Dataset Preparation (5 Points)**
  + Full marks for correctly preparing and using a new dataset.
* **Integration with CLIP (5 Points)**
  + Full marks for accurately integrating the dataset with the CLIP model.

**3. Generating Cosine Similarity Matrix (10 Points)**

* **Correct Matrix Generation (5 Points)**
  + Full marks for generating a 10x10 matrix as specified.
* **Accuracy and Presentation (5 Points)**
  + Full marks for accurately presenting the matrix.

**4. Creating Probability Diagrams (10 Points)**

* **Correct Diagram Creation (5 Points)**
  + Full marks for creating accurate probability diagrams.
* **Clarity and Presentation (5 Points)**
  + Full marks for clear and well-presented diagrams.

**Submission Requirements**

1. **Essay**:
   * Up to 3 pages, Times New Roman, 12-point font.
   * Must follow the structure provided in the project assignment.
2. **Figures**:
   * New matrix of Cosine similarity between texts and image features.
   * Probability diagram for each image of 10 COCO images.

Do not submit the code file to ensure API key security.