Alexandria University
Faculty of Engineering
Computer and Communications
Program

CCE: Pattern Recognition
Assigned: Wed, May 24, 2023

Due: immediately after finals

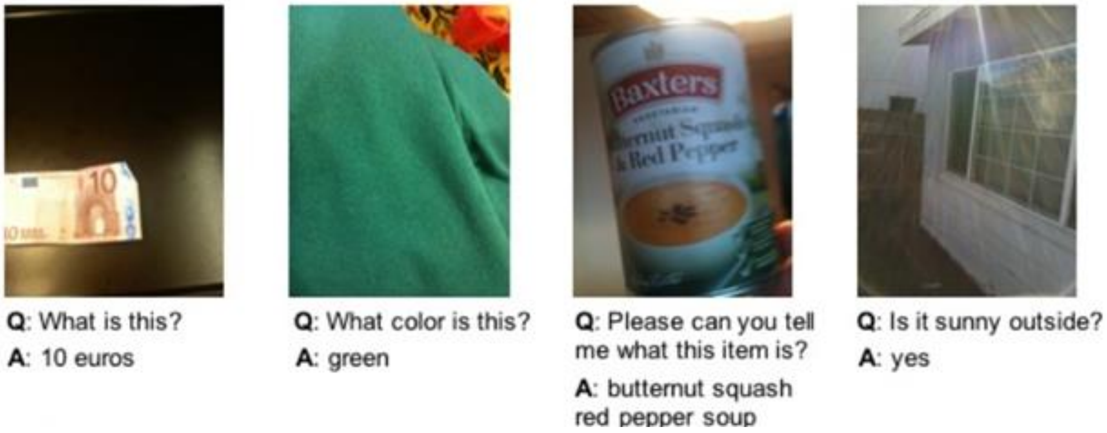
Assignment 4

Visual question answering on (Total 120 Points)

Problem Statement

Visual Question Answering (VQA) is the task of answering open-ended questions based on an image. VQA has many applications: Medical VQA, Education purposes, for surveillance and numerous other applications. In this assignment we will use VizWiz dataset, this dataset was constructed to train models to help visually impaired people.

In the words of creators of VizWiz: “we introduce the visual question answering (VQA) dataset coming from this population, which we call VizWiz-VQA. It originates from a natural visual question answering setting where blind people each took an image and recorded a spoken question about it, together with 10 crowdsourced answers per visual question.”



Q: What is this?
A: 10 euros

Q: What color is this?
A: green

Q: Please can you tell me what this item is?
A: butternut squash red pepper soup

Q: Is it sunny outside?
A: yes

Figure 1: Sample examples from the VizWiz dataset

1. Download the Dataset and Data analysis (20 Points)

a. VizWiz is a VQA dataset that contains 20,500 images/question pairs. Each image has its corresponding question and 10 answers to this question.

   i. 20,523 training image/question pairs
   ii. 205,230 training answer/answer confidence pairs
   iii. 4,319 validation image/question pairs
   iv. 43,190 validation answer/answer confidence pairs

b. The dataset can be found at the following link: https://www.kaggle.com/datasets/ingbiodanielh/vizwiz

c. Take 0.05 of the training data as test data.

d. You will need to analysis the data and show comprehensible histogram of the data.

e. Using Kaggle you will be able to mount the data instantly without any hustle.

f. **Don't forget to set seed 42 and stratify to true and remember that the answer/answer confidence pairs correspond to the images and questions.**

2. Building the Model (60 Points)

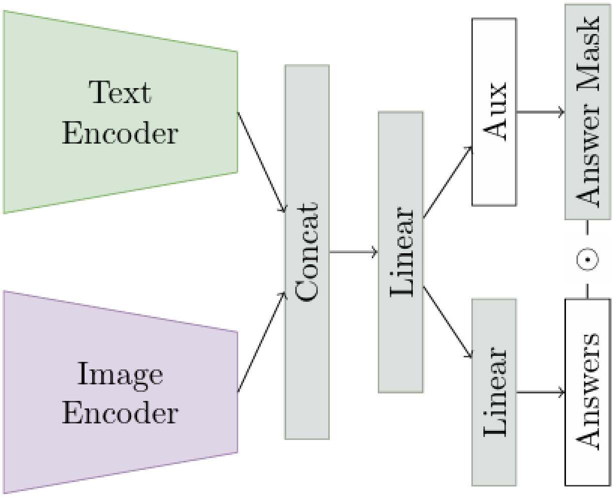


Figure 2: Models architecture

a. You will need to follow the implementation of this paper 2206.05281v1.pdf (arxiv.org)

b. You are allowed to use clip model image encoder and text encoder trained by OpenAi, the following is the paper link and in it you can find the github repo.
2103.00020.pdf(arxiv.org)

c. You will closely follow the implementation in the paper to train the classification layers as stated in the paper: "Our approach utilizes both image and text encoder. The resulting features are concatenated and passed to linear layers with layer normalization and a high dropout value (0.5). As shown in Figure 2. answer types as well as the answers are predicted using an additional linear layer. Image size of the visual encoder is 448x448 for RN50x64 and 336x336 for ViT-L/14@336px. In both cases the linear classifier is trained using cross entropy loss with rotation as image augmentation. We train only the additional linear classifier and use the pre-trained CLIP model as image and text encoder. The CLIP part remains frozen and is not trained on the VizWiz data set, **which allows fast and efficient training without large computational resources."**

d. You must modify the pytorch code provided in the clip repo to implement this model

3. Evaluation (20 points)

- You Should provide all the metrics mentioned in the paper (Accuracy and answerability)
- You need to plot training and valid loss.

4. Bonus (20 Points)

- Write your project report in Latex in the form of a short paper.
- Evaluate your model qualitatively on a set of unseen image-question pairs that we will provide during the discussion.

5. Submission Notes

a. Work in groups of 3 students.

b. **[20 Points]** You are required to submit a clear and detailed report [in PDF format] illustrating every step in the assignment.

Good Luck