Pattern Recognition: Visual Question Answering

Alaa Hossam
Mohammad Helaly
Mohamed Shamarka
Faculty of Engineering
Alexandria University

Abstract— **Visual Question Answering (VQA)**

VQA is a challenging and interdisciplinary research field that combines computer vision and natural language processing (NLP) to enable machines to answer questions about an image. VQA models take as input an image and a natural language question and output an answer in the form of a word or a sequence of words. The goal of VQA is to create machines that can reason about the content of an image and understand the meaning of natural language questions in order to generate accurate and relevant answers.

VQA has numerous real-world applications, such as intelligent personal assistants, autonomous vehicles, and assistive technologies for people with disabilities. VQA is also a challenging research problem that requires the development of novel deep-learning models that can handle the complexity and variability of natural language and visual data. However, VQA still faces several challenges, such as handling rare or out-of-vocabulary words, understanding complex and nuanced questions, and reasoning about the context and relationships between objects in an image.

Overall, VQA is an exciting and rapidly evolving research field that has the potential to transform the way machines interact with and understand visual content. In this project, We train our model in the VizWiz data set and test it in a part of non-seen new images to test its performance.

I. INTRODUCTION

Visual Question Answering (VQA) is a practical application for machine learning that allows computer systems to answer general questions about the content of an image. VQA combines Vision (Images) and Natural Language Processing (Questions about the pictures and answers in the form of text) with high-level reasoning. The goal of VQA is to teach machines to understand the content of an image and answer questions about it in natural language.

VQA applications are important in many fields such as medical fields to detect some diseases in MRI and X-rays as well as VQA is used as an assistant for blind people.
In this project, we VQA data set from Vizwiz contains 20,500 images/question pairs. Each image has its corresponding question and 10 answers to this question. Our model should provide us with the most appropriate answer to each question related to the corresponding image.

The required task is to:

1- Analyze and understand the data set using histograms and different plots, then train the model on the training set and take 0.05 of the training set as a test set to evaluate the model on it.
2- Evaluate the model by plotting the train's loss and validation loss, and providing the answerability and accuracy metrics.

In our implementation, we depend on the official paper of the Vizwiz challenge. We use their model in addition to using Open AI's CLIP model to overcome the problem of exploding gradient in the encoder and PyTorch.

II. Methodology for Visual Question Answering

A. *Preprocessing The data*

We will use the ”Less is More” paper from the VizWiz VQA 2022 challenge to make it easier to build the model using the CLIP model for text and image encoders. but before that w need to preprocess our data by:

a- Reading JSON files to read samples.

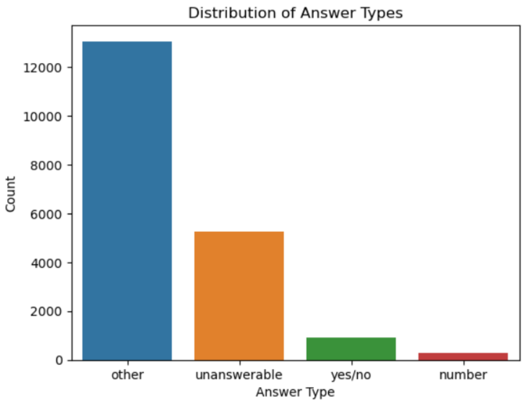
b- Extracting data for each sample by taking the path of each image with its question and the related ten answers, along with its answer type and answerability, then putting the features into data frames.

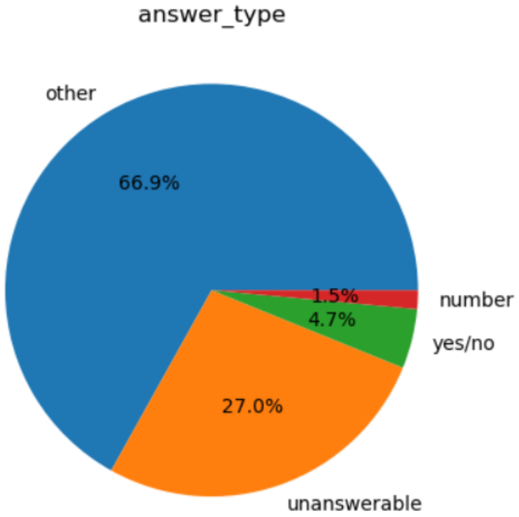
c- Splitting into train and test sets (0.05 of the training set is the test set).

B. *Data Analysis*

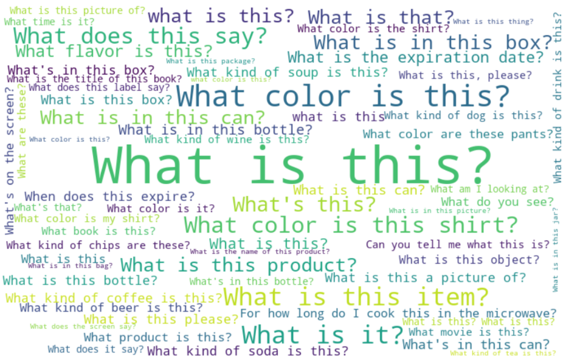
Plotting The data analysis such as histograms, pie charts, word clouds, and tables show the count of each answer type in the data set (numerical, yes/no, other, categorical) and the answerability (Answerable / Unanswerable).

- Plotting pie chart histograms to show the count of each answer type in the data set (numerical, yes/no, other, categorical) and the answerability (Answerable, Unanswerable).

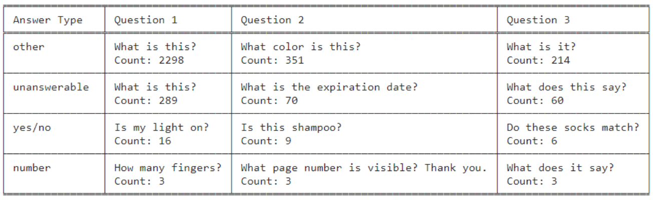




- Plotting word cloud to show the most common questions in the dataset.



- Plotting a table of most frequent questions from different answer types.



Answer Type	Question 1	Question 2	Question 3
other	What is this? Count: 2298	What color is this? Count: 351	What is it? Count: 214
unanswerable	What is this? Count: 289	What is the expiration date? Count: 70	What does this say? Count: 60
yes/no	Is my light on? Count: 16	Is this shampoo? Count: 9	Do these socks match? Count: 6
number	How many fingers? Count: 3	What page number is visible? Thank you. Count: 3	What does it say? Count: 3

All of these plots are made for training, validation, and test sets.

C. *Creating data set class*

Regarding using PyTorch, we need to use the Data set class to load the data in tensors instead of NumPy arrays and process it during the training phase. We will also use this class to load the preprocessed image and question embeddings.

We apply one hot encoding to the answers and answer types, then we make a loop over all data to explore unique answer types

We use tensors to encode the inputs and outputs of our model and the model's parameters. Tensors are similar to NumPy's ndarrays, except that tensors can run on a GPU accelerator to enhance running speed. We need to get the most common answer for each question since we have ten answers for each question.

D. *The Architecture Model*

We use pre-trained image and text encoders from CLIP and train only a simple classification head. CLIP is based on CNN respectively Vision Transformer for image encoding and a Transformer for text encoding. We needn't implement them from scratch.
We are using model ”ViT-L/14@336px”, which is a variant of CLIP based only on vision transformers. CLIP-based architecture that does not require any fine-tuning of the feature extractors.
Layers of the architecture model:

1) A simple linear classifier is used on the concatenated features of the image and text encoder. During training an auxiliary loss is added which operates on the answer types.

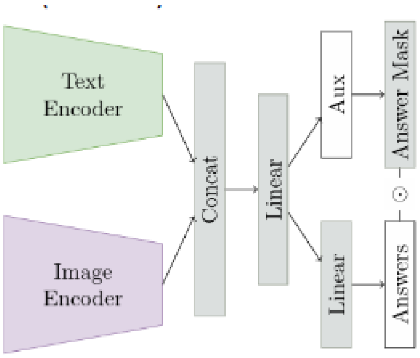
2) The first Linear layer needed is simple: we need to make Layer normalization at first to normalize the data coming from the concatenated image and text features, then make a dropout (0.5) to avoid over-fitting, at the end of this layer we initialize a linear transformation module with a specified input size (The projection matrix size of the text features and image features) and output size.

3) The second Linear layer: we need to make Layer normalization at first to normalize the data coming from the previous linear layer, then make a dropout (0.5) to avoid over-fitting, at the end of this layer we initialize a linear transformation module with a specified input size (the output of the previous hidden layer) and output size (i.e. the number of classes which it the number of answers).

4) The auxiliary layer: We introduce an auxiliary loss in this layer for answer type prediction. This loss helps to learn an answer masking for the eight answer types other, numbers, yes/no, unsuitable/unanswerable.so the output of this layer is the predicted answer type.

5) The last step in the model is to multiply the output of the second linear layer by the output of the auxiliary layer to produce the final predicted answer.

6) Note that we had to modify the architecture in the diagram by adding an extra linear and sigmoid layer for answerability prediction.

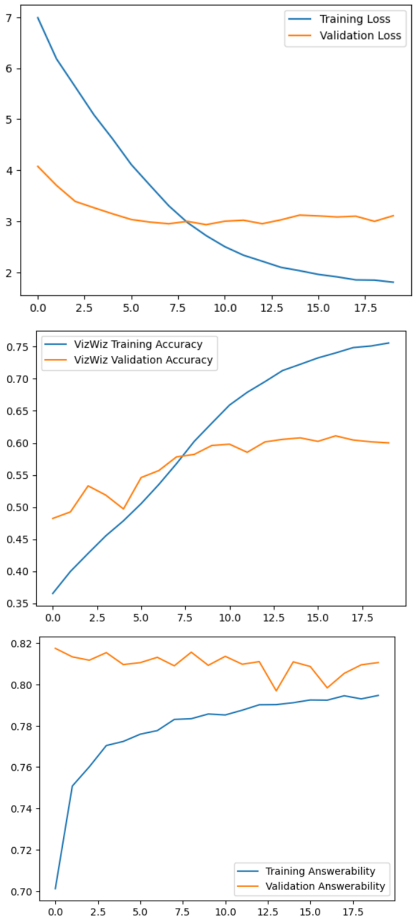


E. Evaluation of the model

Training the model and Evaluating the model the answerability and accuracy metrics. In the accuracy metric; we have used the role mentioned on the VizWiz VQA Challenge 2022 webpage which is:

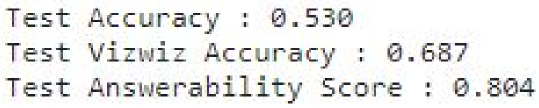
- Accuracy = min (1, (number of people provided the same answer/3)) We have used some torch functions to calculate the number of people who provided the same answer since we need element-wise comparison here as well as we need Boolean tensor and the sum of the vectors.

- Regarding the answerability, we consider it as a new classification in the form of predicting whether the question is answerable or unanswerable, and then we determined the accuracy of this classification.



- After rigorous testing and hyperparameter tuning, we came to the conclusion that our model functions best with batch size = 32, learning rate = 0.0005 and that the model plateaus in improvement after 20 epochs.

- The highest accuracy we achieved was 68 percent and the highest answerability we achieved was 80 percent.



```
Test Accuracy : 0.530
Test Vizwiz Accuracy : 0.687
Test Answerability Score : 0.804
```

III. CONCLUSIONS

In conclusion, Visual Question Answering (VQA) is a complex and challenging research field that combines computer vision and natural language processing to enable machines to answer questions about an image. VQA has numerous real-world applications and has the potential to transform the way machines interact with visual content.

Recent advances in deep learning, particularly the development of pre-trained language models, have significantly improved the performance of VQA models. However, VQA still faces several challenges, such as handling rare or out-of-vocabulary words, understanding complex questions, and reasoning about the context and relationships between objects in an image.

Despite these challenges, VQA is an exciting and rapidly evolving research field that holds great promise for the future of AI. Continued research and development in VQA will lead to further improvements in the accuracy and performance of VQA models and ultimately enable machines to better understand and interact with the visual world.

References

[1] “Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[2] “Danna Gurari, Qing Li, Chi Lin, Yinan Zhao, Anhong Guo, Abigale J. Stangl, and Jeffrey P. Bigham. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.