Visual Question Answering
Using Transformers

Problem

In this project, We will develop a VQA-based smart personal assistant that can answer users' questions about their daily life activities. The system will be able to recognize images and answer questions about them, such as "What is the color of the shirt in the image?" or "What is the brand of the car in the image?".

The user will upload an image and enter any questions about that image. The model will process this image and the question and return the answer.

Motivation

The development of a VQA-based smart personal assistant can have a significant impact on our daily lives by providing a more natural and intuitive way to interact with technology. Such systems can be used in various applications, including smart homes, healthcare, and education. Moreover, the techniques used in this project can be extended to other AI applications, such as chatbots and autonomous systems.

what makes it difficult:
the difficulty of finding suitable data, and applying preprocessing to make it work with transformers from (hugging face)

Related work

*https://arxiv.org/pdf/1511.05099.pdf
Yin and Yang: Balancing and Answering Binary Visual Questions (CVPR 2016)
it just answers with yes or no and doesn't give us any details.

* https://arxiv.org/abs/1707.07998
"Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering" by Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. This paper presents a state-of-the-art model for visual question answering that uses both bottom-up and top-down attention mechanisms.

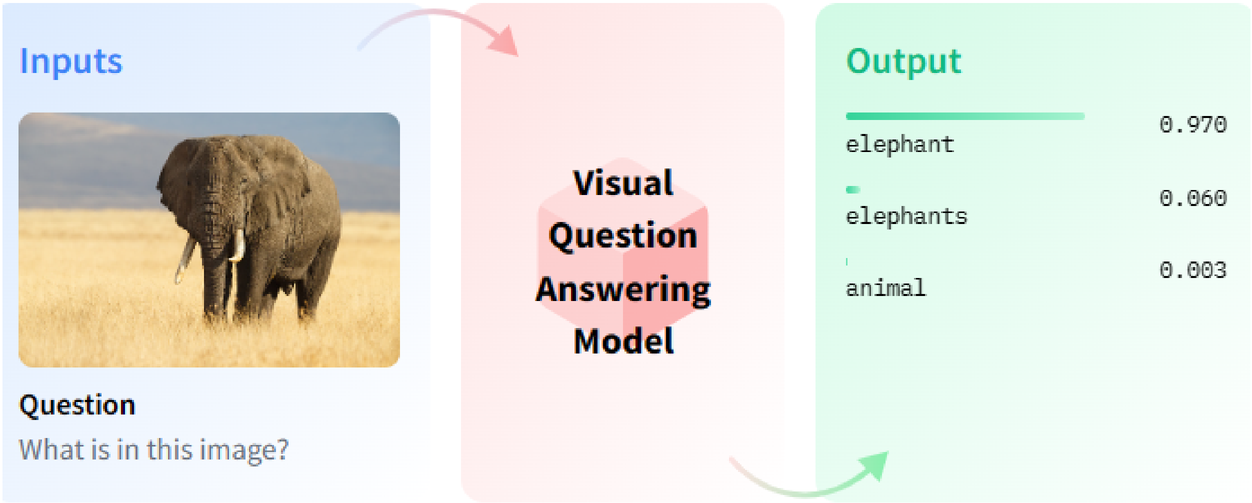
* https://arxiv.org/abs/1706.03762
"Attention Is All You Need" by Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. This paper presents the Transformer model, which has become a standard model for natural language processing tasks and has been used in various VQA models.

Algorithm

We decided to use the ViltForQuestionAnswering algorithm to build our project, there is an alternative for doing the same function like:
- VisualBertForQuestionAnswering

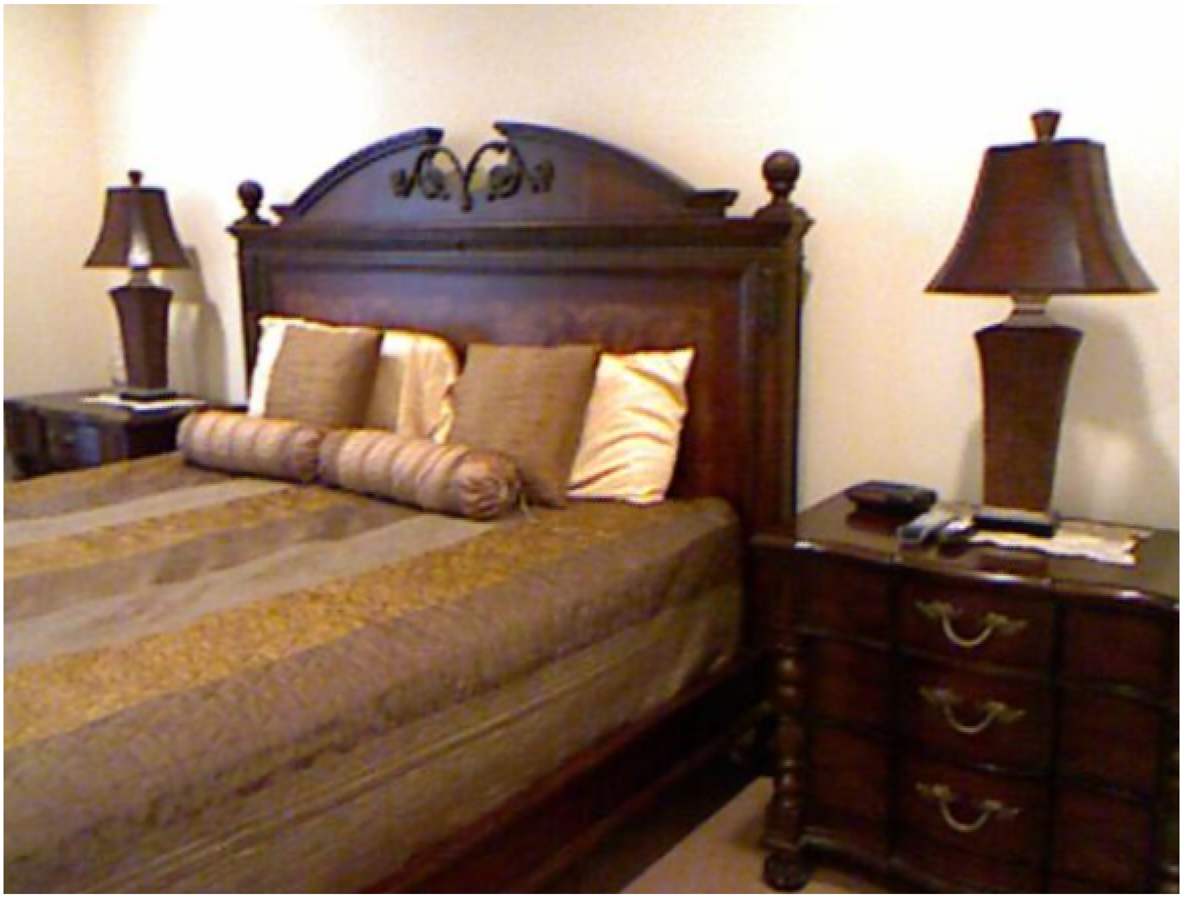
We used ViltForQuestionAnswering as it required less space for data than VisualBertForQuestionAnswering so it will be easy to be trained and faster

Result



Question
What is in this image?

```
elephant     0.970
elephants    0.060
animal       0.003
```



```
Question:       what is the colour of the lamps
Answer:         brown  (Label:76)
```

Evaluation

For the evaluation of the model, we tested it and it works with good accuracy, the input format is:
- image
- question

We enter these inputs and get the answer as a result, we decided to use the algorithm of this model to retrain it with a new dataset to refresh the labels of the features of the model, after doing a preprocessing on the dataset, we tried to train the model but we face some issues and failed at the first trial to build the model, we still trying to fix these issues to complete the evaluation stage.

Analysis

Initially, our team achieved high accuracy with our first approach, which prompted us to train our transformer from scratch. However, we encountered several issues during this process.

Firstly, we opted to use a transformer from Hugging Face and sought to use datasets from the same website for faster computation and processing. Unfortunately, the size of the data proved to be too much, causing the program to crash repeatedly.

Subsequently, we tried obtaining the datasets from Kaggle, but this approach led to slower computation times since the data was from a different website.

Despite these setbacks, we continued with our second approach. Unfortunately, we encountered further issues due to the large dimensions of the input, which resulted in program crashes. Additionally, the complex computations required for this approach exceeded our available resources, further impeding our progress.

Contribution division

Ahmed Kamal & Ibrahim Hamdy
Searching for data and working on preprocessing of it

Rahma Ashraf & Reem Abo Bakr

Working on choosing suitable transformers like Bert tokenizer & Vilt feature extractor

Mohamed Khaled & Moaz Atef

Working on combining images with their questions using Vilt feature extractor and Bert tokenizer

Antonios Malak & Habiba Ahmed

Working on combining all the above with training and improving accuracy