MARMARA UNIVERSITY
FACULTY OF ENGINEERING

Visual Dialog Application for Visually Impaired Individuals

Tuncer Cem Uğurluer, Mustafa Talha Öztürk

GRADUATION PROJECT REPORT

Department of Electrical and Electronics Engineering

Supervisor

Prof. Dr. Cabir Vural

ISTANBUL,2023

MARMARA UNIVERSITY
FACULTY OF ENGINEERING

Visual Dialog Application for Visually Impaired Individuals

by

Tuncer Cem Uğurluer, Mustafa Talha Öztürk

June 6, 2023, Istanbul

SUMBITTED TO THE DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE

OF

BACHELOR OF SCIENCE

AT

MARMARA UNIVERSITY

Signature of Author(s) ……

Department of Electrical and Electronics Engineering

Certified By ……

Project Supervisor, Department of Electrical and Electronics Engineering

Accepted By ……

Head of the Department of Electrical and Electronics Engineering

ACKNOWLEDGEMENTS

I would like to extend my heartfelt gratitude to Prof. Dr. Cabir Vural, who is my supervisor, for his help and support.

I am also grateful to the Kadıköy Association for the Visually Impaired's assistance, which ensured the study was on point.

Finally, I would like to thank the Hugging Face and Open AI teams for offering pre-trained models and APIs that were crucial for this study.

June, 2023

Tuncer Cem Uğurluer, Mustafa Talha Öztürk

TABLE OF CONTENTS

ABSTRACT

This study presents the development and evaluation of a visual dialog application designed to help visually impaired individuals in understanding the visual world that we all live in. The goal of this work was to build a system that can answer questions about images, generating captions for these images, and can generate dialog like a human.

The work began with the creation of a unique dataset that contains unique indoor images, each with ten questions and ten corresponding answers. The dataset was generated in line with the feedback of the Kadıköy Association for the Visually Impaired, images from DAQUAR, MIT Indoor and House Rooms Image datasets. The dataset served as a main component to train the neural network architectures.

Various model training methods were explored, including the Memory Network Encoder, Hierarchical Recurrent Encoder, and Late Fusion Encoder as mentioned in the “Visual Dialog” article. Although the number of images and question answer pairs was raised, the models training results were not good enough, with accuracy only reaching %8.

Against the troubles we face, a pre trained model, ViLT (Vision-and-Language Transformer Without convolution or Region Supervision), was deployed and fine-tuned using the dataset created by us. This approach improved the performance of the system.

The final product was developed using fine-tuned ViLT, GPT-3 API, and the vit-gpt2-image-captioning model. The system works by inputting an image and a question into the VQA model. The same image is inputted into the image-captioning model. After descriptive caption was generated these two string outputs, together with a carefully crafted prompt are fed into the GPT-3 model, which then generates a human-like dialog response about the image to the user.

The software makes an important step in creating technology for visually impaired individuals, providing them with a tool to gain information about their surroundings through interactive dialogue with images. Our future work will be expanding the capabilities and accessibility of the application.

LIST OF SYMBOLS

$\boldsymbol{w}$: word

$\boldsymbol{V}$: vocabulary

$\boldsymbol{d}$: dimension

$\boldsymbol{I}$: Image

$\boldsymbol{n}$: patches

$\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$: query, key, and value matrices respectively

$\boldsymbol{d_k}$: dimensionality of query

$\boldsymbol{P}$: set of all patches

$\boldsymbol{e}$: patch position pair

$\boldsymbol{F}$: filter

$\boldsymbol{O}$: output matrix

$\boldsymbol{W_1}, \boldsymbol{W_2}$: weight matrices of the fusion and final layers

$\boldsymbol{b_1}, \boldsymbol{b_2}$: bias vectors of the fusion and final layers

$\boldsymbol{sigma_1}, \boldsymbol{sigma_2}$: activation functions

ABBREVIATIONS

API: Application Programming Interface

DAQUAR: Dataset for Question Answering on Real-world Images

ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision

GPT: Chat Generative Pre-Trained Transformer

CV: Computer Vision

VQA: Visual Question Answering

VD: Visual Dialog

IC: Image Captioning

DQA: Document Question Answering

GPU: Graphics Processing Unit

IEEE: Institute of Electrical and Electronics Engineers

ISO: International Organization for Standardization

IEC: International Electrotechnical Commission

PII: Personal Identifiable Information

AI: Artificial Intelligence

ViT: Vision Transformer

NLP: Natural Language Processing

YOLO: You Only Look Once

RGB: Red, Green, Blue

Wups: Wu-Palmer Similarity

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Visual impairments significantly inhibit an individual's ability to understand and interact with their environment. This occasion, which prevents access to visual information, makes it very difficult to live and learn in our world where most of the information is visual. Therefore, it becomes so important to develop tools and software's that can bridge the information gap and help visually impaired individuals to understand the information in their surroundings. This work aim is to solve the addressed problem by developing a visual dialog application capable of understanding visual content and interacting with user in a human-like dialog-based format.

Understanding visual format can easily be done for healthy humans but a complex task for machines to replicate. The field of CV, a subfield of image processing and artificial intelligence tries to mimic this human ability, with the final aim of enabling machines to understand, interpret visual information as humans do. Though the huge advances in the field, achieving this goal is nearly impossible for now due to the complex and abstract nature of visual understanding.

Our main motivation in this thesis is to build a system that can process visual data into a form that can be understandable for visually impaired individuals. The system takes image as an input, answer questions about the image, generates a caption and detects objects in the image. After these steps are done the system uses these elements to generate human-like dialogs about the image. This approach not only not only provides information about image to visually impaired individuals, but it can also answer any question they want with the power of GPT-3.

The secondary motivation is to explore pre-trained models for overcoming the challenging training process of deep learning models for visual information understanding. There are lots of training techniques, but it is difficult to achieve due to the huge amount of data and computational resources required for the task. Our method attempts to solve this issue by utilizing pre-trained models, fine tuning them with our unique dataset developed for this purpose and evaluating its performance.

Lastly, this work also tries to integrate different deep learning models to create a more comprehensive system. By combining a VQA model, image-captioning model, object classification model and a dialog generation model, this approach brings together different aspects of AI, with goal of creating a more robust and a better system for visually impaired individuals.

In summary, this work tries to solve problem of visual information access for visually impaired individuals, explores the use of pre-trained models, and investigates the integration of different models to create a more comprehensive system.

1.1. Thesis Content

2.Research Objective: Aim of the work and objectives.

3. Related Literature: Information and discussion of articles related to this study.

4. Design: Design procedures of this project such as cost, engineering standards.

5. Methods: Detailed explanation of methods in this study. Deep dive into algorithms, mathematics, and physics behind this project.

6. Results and Discussion: Examining the results of the project, comparing the models, discussing the shortcomings of the study and how it will develop.

7. Conclusion: Summary of the work done in this study.

2. RESEARCH OBJECTIVE

The primary objective of this work is to develop a visual dialog application that can engage in a dialog about images with visually impaired individuals. This study focuses on addressing the limitations posed by visual impairment and attempts to bridge the gap between visually impaired individuals and the visual world. The objectives of this work are:

2.1. Dataset Creation:

To develop a unique dataset consisting of interior images, each image has ten questions and corresponding ten answers. The creation of this dataset is based on the feedbacks from Kadıköy Association of Visually Impaired. The dataset used for training and fine tuning the models that are in this project.

2.2. Model Training and Evaluation:

To train a deep learning model on our newly created dataset and various encoding techniques, this process involves exploring different model training techniques and determining their effectiveness with evaluation metrics.

2.3. Model Fine-tuning:

To fine tune a pre-trained model with the newly created dataset. This objective addresses the challenge of good results with initial model training and aims to improve model's performance by well-trained models on huge datasets.

2.4. Model Integration:

To integrate a VQA model, IC model, object detection model and a dialog generation model to develop a complex and comprehensive visual dialog system.

2.5. System Evaluation:

To evaluate the performance of the developed VD system and determine its effectiveness. This involves testing the system with real user feedback and some metrics.

By achieving these objectives, this study aims to make a visual dialog system that can generate human-like dialogs about the information in the image.

3. RELATED LITERATURE

The field of artificial intelligence has seen important advancements through years in various domains, several of these domains is highly related to our work. The current study interacts with multiple research themes, including Visual Question Answering (VQA), image captioning, object detection, language models, and Document Question Answering (DQA).

The concept of VQA, where models can answer text-based questions about images, has seen enormous attention in AI research. For example, Antol et al., 2015, and Zhou et al., 2020, made great advancements in this field but their work primarily focused on answering questions with single answers such as “what is in the box!” and “bottle” but our aim is to create conversations with the user.

Like VQA, image captioning is another significant domain of AI where descriptive textual information about images are generated. Important contributions in this field have been made by Vinyals et al.

Research in the field of object detection has also seen extensive development, with model that can detect and classify objects in an image. For example, Huang et al., 2016, developed modern CNN object detection algorithms that can trade-off between accuracy and speed.

With the invention of transformer-based language models like GPT-3 (Brown et al., 2020) revolutionized the generation of human-like text, demonstrated high accuracy in language tasks. But these models don’t have the ability to understand visual information.

Moreover, the most important research for our study is in the visual dialog systems, Das et al., 2017, presented a dataset and variety of different techniques and models for visual dialog tasks. But their work requires huge amount of data.

In this context, our work integrates the capabilities of these different areas and bring them together to create a system that can understand images, answer questions about images and generates human-like dialog about images or any related question.

4. DESIGN

4.1. Realistic constraints and conditions

Designing a visual dialog system for the visually impaired consist of numerous and serious constraints and conditions.

4.1.1. Environmental and Sustainability Issues:

As a software-based solution, our system does not have direct environmental impact. However, the energy consumption of training and running neural networks can be a problem. Cautions must be taken to ensure the model trainings are efficient and doesn't use unnecessary computational resources.

4.1.2. Manufacturability:

Since our system is completely software-based, the concept of manufacturability relates to deployment and implementation. The system must be compatible and optimized for different platforms and devices, providing accessibility for all the users with different devices.

4.1.3. Ethics:

Ethical considerations are crucial for AI systems that collects and works with huge amounts of data. The system must respect to user privacy, not storing personal data and system must be cleaned from potential biases to ensure fair representation of information.

4.1.4. Health and Safety:

The health and safety considerations of our system revolve around usability and the accuracy of the given information. The system should be designed to prevent misinformation that could lead to dangerous situations.

4.2. Cost of the design

The primary cost is the time required to create dataset.

Another important cost is the cost of training of the VQA and Visual Dialog systems. This cost includes hardware costs such as GPU and cloud-based coding platform Google Colab ($35/month)

Another significant cost is the cost of Open AI's GPT-3 API ($0.02 / 1k tokens)

Table 1. Cost of the design

	GPT-3 API	Google Colab	Google Drive
COST:	$0.02/1k tokens	$35/month	$9/month

4.3. Engineering Standards

The engineering standards used in the design of our system are important to ensure compatibility and accessibility.

4.3.1. Software Standards:

The software development process is responsible for IEEE Standard 730-2014. This standard provides guidelines for planned, systematic and important approaches to software quality assurance.

For security of coding, we follow the guidelines outlined in ISO/IEC 27034. This standard provides a useful framework for assuring security in application services, it is crucial for systems like our that dealing with sensitive user data.

4.3.2. Data Standards:

Our system deals with image data and user-generated text data, we obey the ISO/IEC 27018:2019. The code of practice for protection of personally identifiable information (PII) in public clouds. This standard ensures that safety for storage of personal data.

4.3.3. AI Ethics and Bias Standards:

Our work obeys to guidelines such as IEEE P7003- Standard for Algorithmic Bias Considerations. This help identify and destroy bias in AI systems, ensuring that system treats all users fairly and evenly.

4.4. Details of the design

The purpose of our design is to build a visual dialog system that assists visually impaired individuals by enabling them to understand the visual world around them. The system is built using state-of-the-art AI models for image understanding and text generation. The consist of five connected modules, each model is responsible for a specific purpose. You can see the descriptions of the models below.

4.4.1. ViLT for VQA:

This VQA model uses the Vision-and-Language Transformer (ViLT) to answer various questions about images. This understanding and information after used to respond to questions about the images in a dialog form.

4.4.2. Image Captioning Module:

The image captioning module uses vit-gpt2-image-captioning to generate a detailed caption about the provided image, then the caption is fed to the GPT-3.

4.4.3. Document Question Answering:

DQA is a task in Natural Language Processing (NLP) where a system is given a document or set of documents. Purpose of the DQA module is to answer questions about related documents.

For DQA task we used pre-trained layoutlm-document-qa model the model trained on SQuAD2.0 and DocVQA datasets.

4.4.4. Object Detection:

Object detection is a CV task that identifies object in images or even videos. It’s a two-step process, first step is to determine where the objects are (localization) and determining what

they are (classification). We used You Only Look Once (YOLO) algorithms to detect the objects and then fed the information to the GPT3-API to increase the accuracy of the answers.

4.4.5. Natural Dialog Generation:

Natural dialog generation refers to the process of being able to create human-like conversations using machines. To accomplish this task, we used Generative Pretrained Transformer 3 (GPT-3). GPT-3 is a state-of-the-art powerful language model developed by OpenAI that is good for dialogue generation. In our design all the outputs of previous models fed into the GPT-3 API to generate a human-like accurate answer to questions about images in dialogue format.

5. METHODS

In this section, we will dive into details about our design and implementations of the visual dialog module which is integrations of five different models. The system is designed to take an image and a question as input. The output of the system is the answer about the question in a human-like dialogue. The module is implemented using TensorFlow, PyTorch, Pandas, Numpy and relies on pre-trained models that we explain in detail below.

5.1. Dataset

The dataset we used in this study is a combination of DAQUAR, MIT indoor and House Room Images dataset. To write the question answer pairs we used the help of our friends and family.

	question	answer	image_id
0	what is on the right side of the black telepho...	desk	image3
1	what is in front of the white door on the left...	telephone	image3
2	what is on the desk	book, scissor, papers, tape_dispenser	image3
3	what is the largest brown objects	carton	image3
4	what color is the chair in front of the white ...	red	image3

Figure 1. General View of The Dataset

The final version of the dataset includes 4234 pictures and 84680 questions and answers. For our study to be suitable for its purpose, we received feedbacks from the Kadıköy Association for the Visually Impaired, so the questions we prepared became more suitable for our purpose.

Figure 2. Most Used Words in Our Dataset

Before inserting the images in the dataset, we fixed the resolution of the images to 224x224 and normalized the RGB values. In this way we speed up the train process and reduced the load on our hardware.

```
{
    "image": "image1.jpg",
    "question": "What color is the chair in the picture?",
    "answer": "red"
}
```

Figure 3. JSON Data Format

After preprocessing our images and set our dataset we converted our data to JavaScript Object Notation (JSON) to make it easier to store. In the JSON format data structure is enclosed in curly braces “{}” to indicate that it is an object, each piece of data in the object is a in the form of key-value pair, such as “image”: “image1.jpg”.

5.2. ViLT for Visual Question Answering:

ViLT is a pre-trained architecture for multimodal vision-and-language understanding tasks. It incorporates visual embeddings with language embedding to understand the patterns and relationship between two different domains.

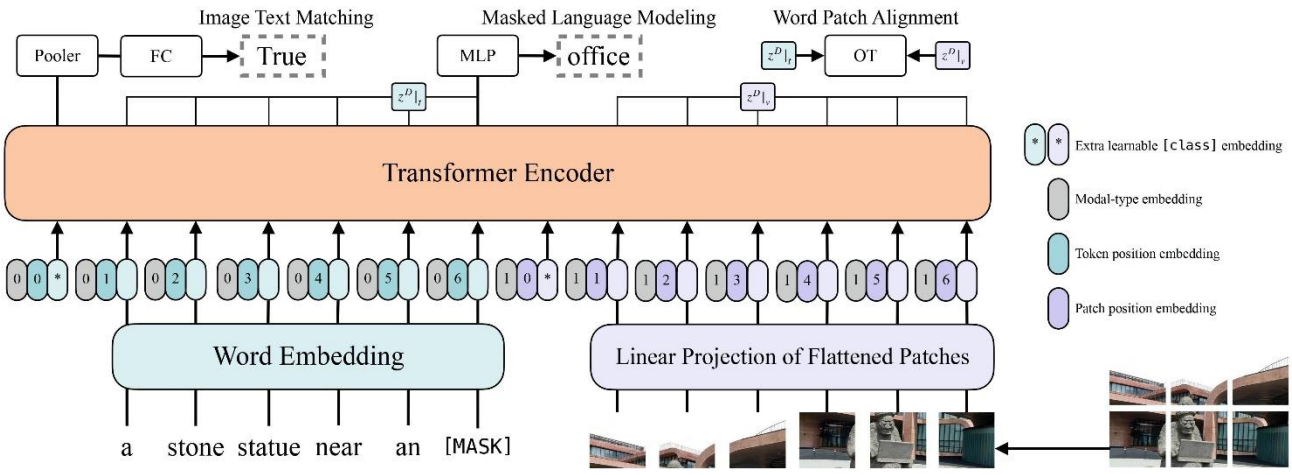

Figure 4. ViLT Architecture

5.2.1. Word Embeddings:

For the first step model transforms each word "w" from vocabulary "V" into a "d" dimensional vector with the help of mapping function "$E: V -> R^d$", this helps the model to understand the meanings of the words.

5.2.2. Linear Projection of flattened patches:

The model then divides an image "I" into "n" patches. Each patch represented as "$p_i(for\ i\ in\ \{1,..,n\})$", after this step is competed then each patch transformed into "d" dimensional vector space with the help of linear transformation "$L: R^m -> R^d$", which results to "$v_i = L(p_i)$". This transformation enables the model to grasp visual data in a way that it is consistent with its comprehension of words.

5.2.3. Transformer Encoder:

This is the part where model gain its understanding of both visual and language inputs.

For given query, key, and value matrices "Q, K, V" and the dimensionality "d_k" of the query and keys, the output of self-attention layers can be defined as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{sqrt(d_k)}\right) * V.$$

The encoder applies a self-attention mechanism to identify the most prominent parts of the image and caption by doing this it gains a better understanding for image and language relationship.

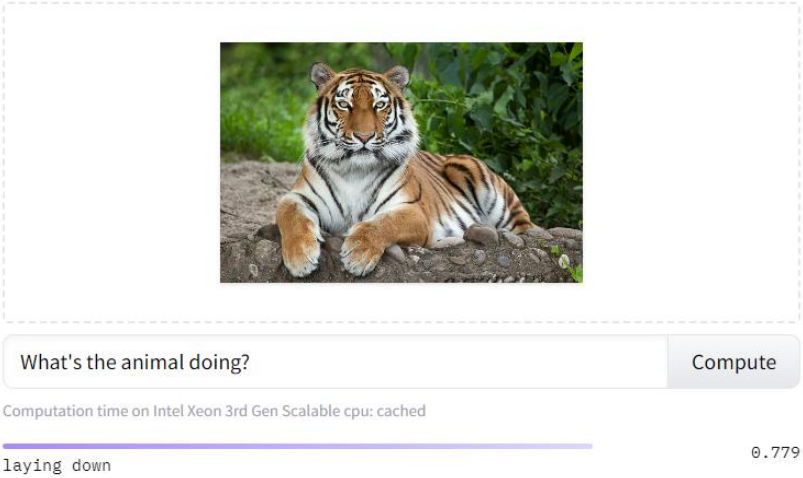

Figure 5. ViLT Visual Question Answering Model Prediction

5.3. ViT for Image Captioning:

ViT uses a transformer architecture to handle image-based tasks. The goal is to generate captions that can accurately describe the images.

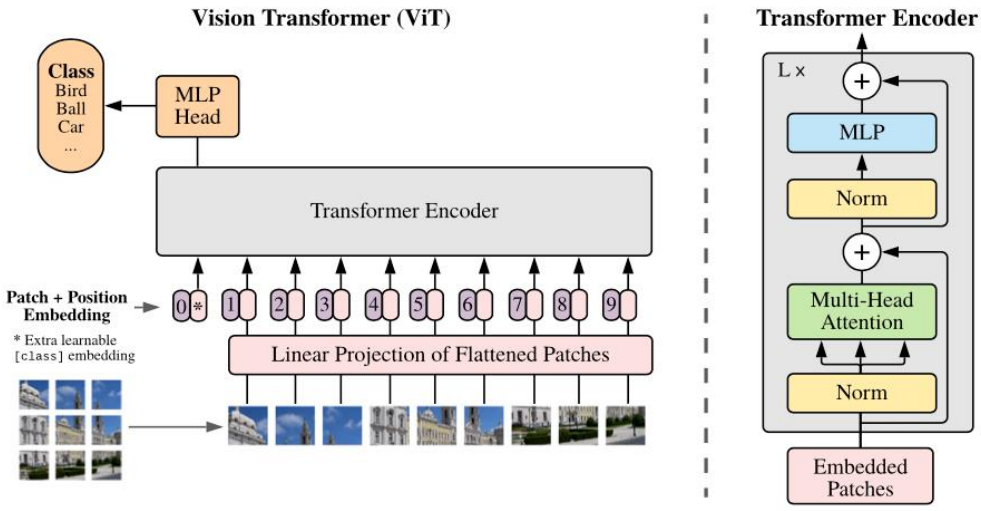

Figure 6. ViT Architecture

5.3.1. Patch-position Embedding:

For the first step the model breaks down the input image into patches then applies linear projection to them and adds positional encodings. If "P" is the set of all patches, "E" is the embedding function mapping each patch to a "d" dimensional vector space, and "pos" is a function that provides a positional encoding for each patch, then for a given patch "p" the embedded patch-position pair "e" is given by:

$$e = E(p) + pos(p).$$

Figure 7. ViT Image Captioning Model Prediction

5.4. YOLO for Object Detection:

YOLO is a real-time object detection system. Working principle of YOLO is it first frames object detection as a regression problem to spatially separated bounding boxes and associated class probabilities.

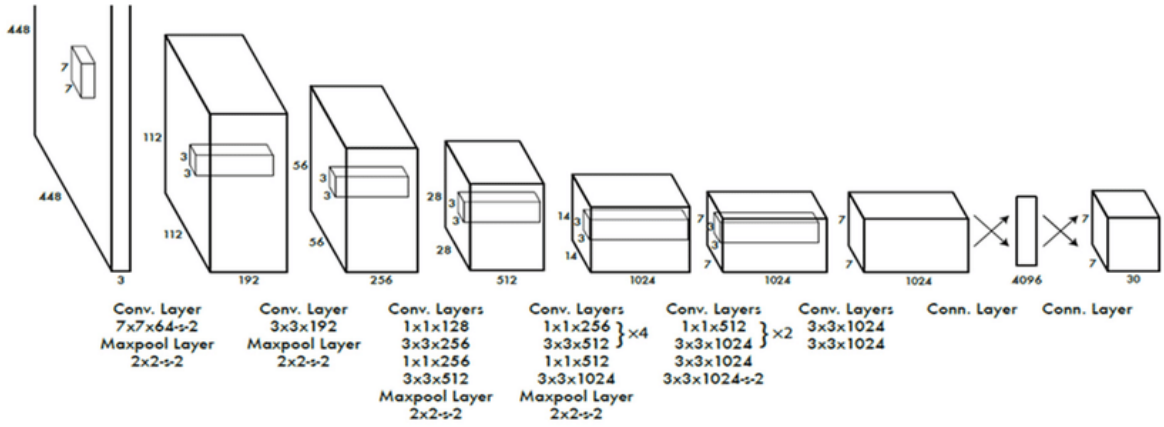

Figure 8. YOLO Architecture

5.4.1. Convolutional Layer:

YOLO uses convolutional layers to extract information from images. Given an image input matrix "I" and a filter "F", the output matrix "O" is obtained by applying the filter over the image input matrix, which can be represented mathematically as:

$$O[i,j] = \ sum\ for\ all\ m\ in\ \{-a,..,a\}, n\ in\ \{-b,..,b\}(I[i+m, j+n] * \ F[a+m, b+n]),$$

Where " $*$ "denotes convolution operation.

5.4.2. Max Pooling Layer:

The main usage of the max pooling layer is to reduce the spatial dimensions (width and height) of the input. It scans the input with a window of size "$k\ x\ k$", and for each window, it outputs the maximum value inside the boundaries in that window. The operation can be represented as:

$$O[i,j] = \max for\ all\ m\ in\ \{0,..,k-1\}, n\ in\ \{0,..,k-1\}(I[s*i+m, s*j+n]).$$

Where "I" is the input, "O" is the output and "s" is the stride.

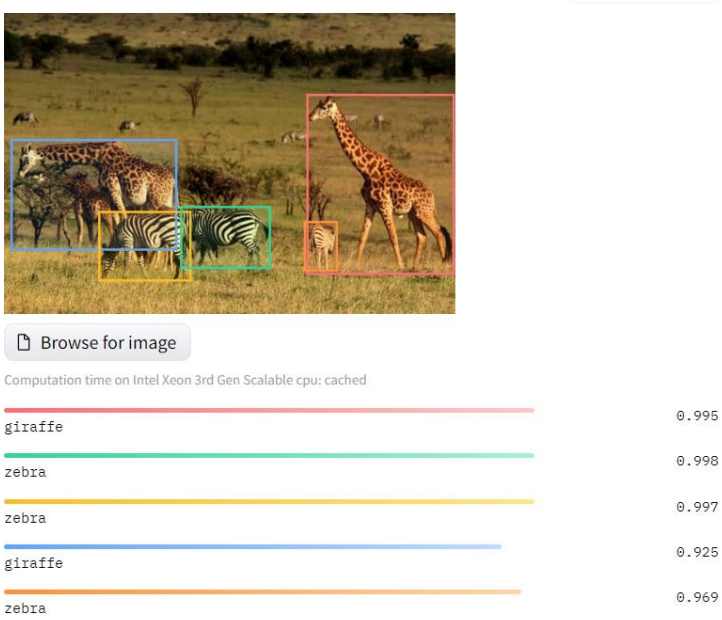

Figure 9. YOLO Object Detection Predictions

5.5. GPT-3 for Dialog Generation:

GPT-3 is a state-of-the-art language model that uses transformer architecture to generate highly coherent and contextually relevant text.

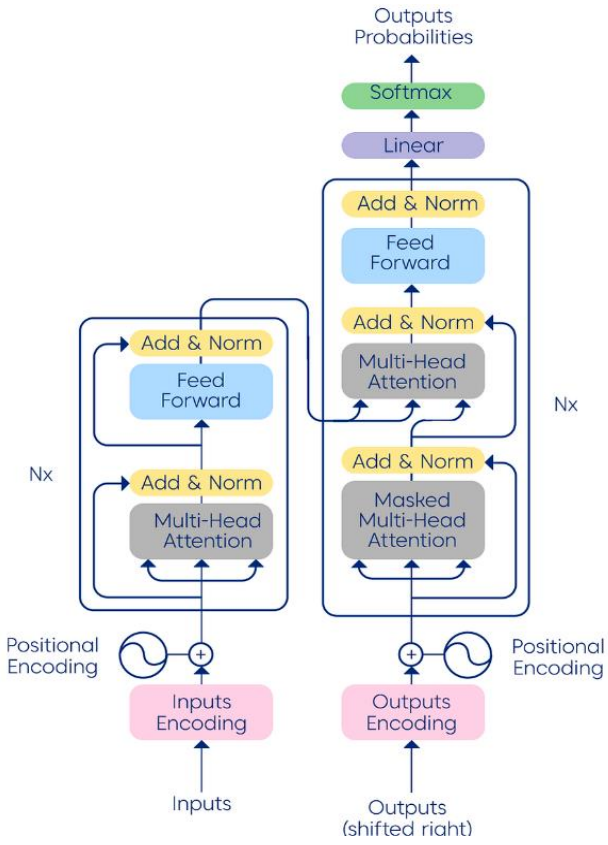

Figure 10. GPT-3 Model Architecture

5.5.1. Multihead Attention:

The Multihead mechanism allows GPT-3 to focus on different parts of the input sequence for the attention head. For each attention head "$i\ (for\ i\ in\ \{1, \dots, h\})$", the multihead attention can be computed as:

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right),$$

$$MultiHead(Q, K, V) = Concat(head_1, \dots, head_h)W^O.$$

5.5.2. Feedforward Neural Networks:

The Feedforward Neural Networks provide the model to learn much more complex representations by introducing non-linearity. Given an input "x", weights "W_1, W_2", biases "b_1, b_2" , and activation functions "$sigma_1, sigma_2$", the output "$F(x)$" of a simple feedforward neural network with only one hidden layer is :

$$F(x) = sigma_{2\left(W_2 * sigma_{1\left(W_1 * x + b_1\right)} + b_2\right)}$$

5.5.3. Masked Multihead Attention:

The Masked Multihead Attention is almost same as Multihead Attention, but it ensures that the prediction for position "i"can only depend on the known outputs at positions less than "i".

This is crucial for tasks like language generation where the model must predict one token at a time.

5.6. Visual Dialog Module:

The Visual Dialog model we created for our study is provided by the integrated operation of four different models. The VQA model takes an image and related question as input and answers the question with a single word. The image captioning model takes the image as input and gives a detailed description of the image, the object detection model takes the image as input and outputs all the object it sees in the picture. Then, the outputs of the 3 models mentioned are fed as inputs to the GPT-3 API with a clever prompt. GPT takes on the task of

creating dialogue from one-word answers. You can see the flow chart of the module mentioned below:

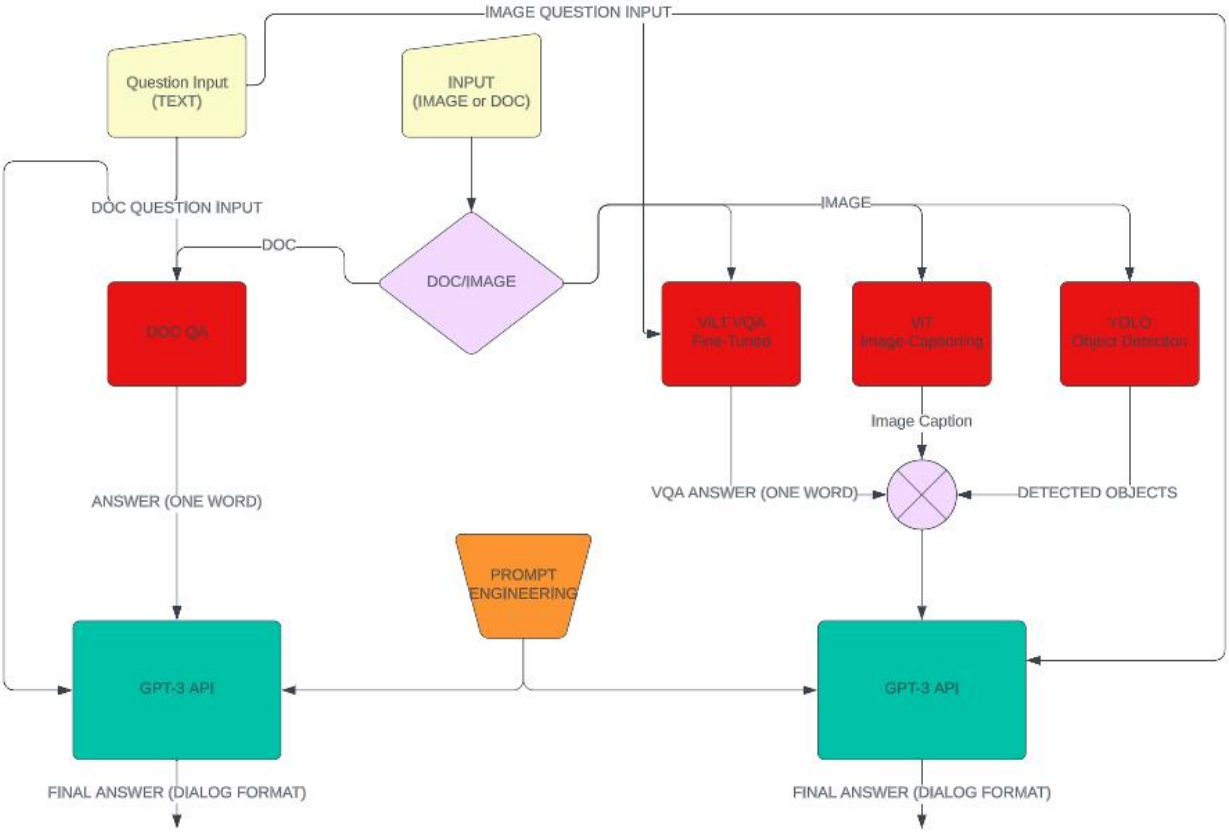

Figure 11. Visual Dialog Module Flow Chart

6. RESULTS AND DISCUSSION

We started our work by trying the architecture in the original Visual Dialog paper, although we repeatedly train the models with suggested encoder architectures in the “Visual Dialog” paper we cannot achieve accurate predictions. Visual Dialog task is very data dependent and requires huge amounts of data. 140.000 pictures and around 14 million questions and answers were used in the original paper.

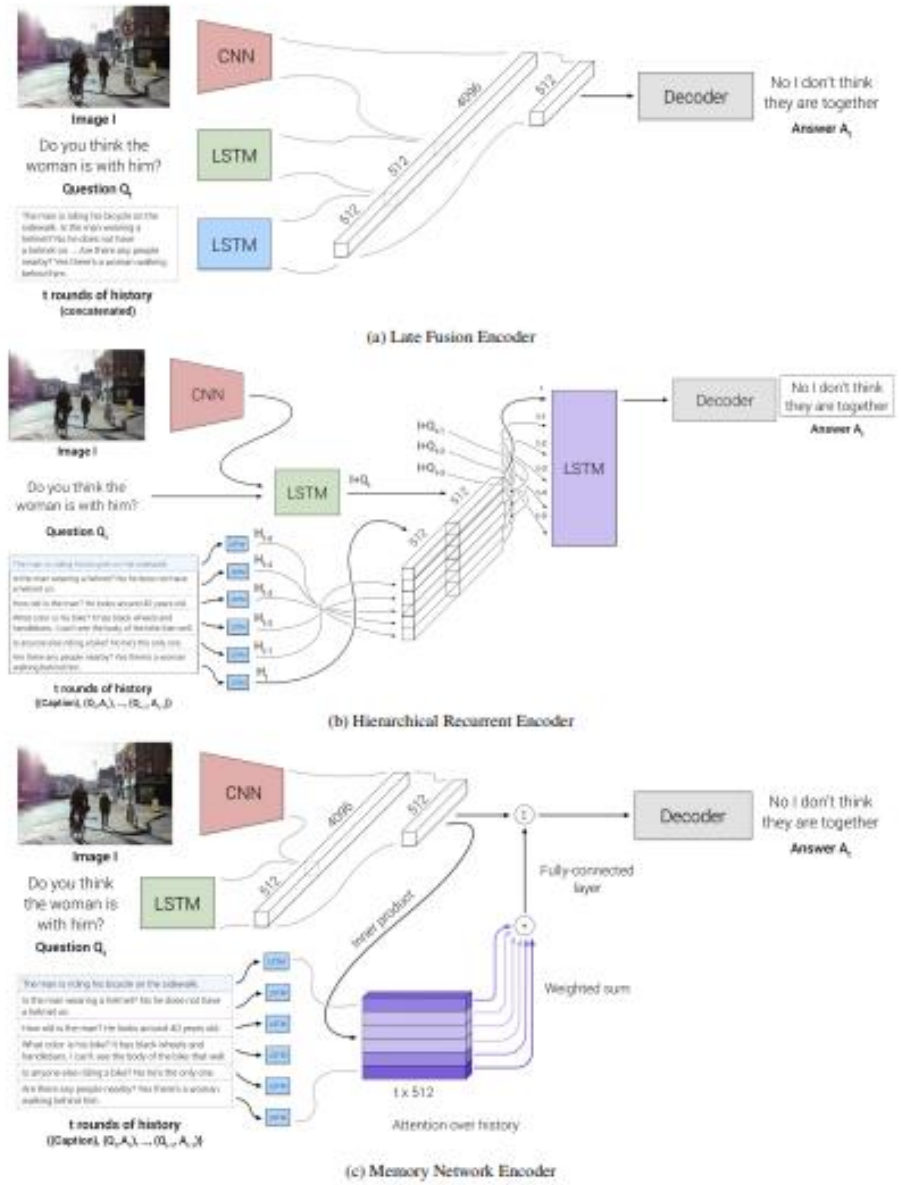

Figure 12. Encoders

When the results did not go the way, we wanted and it was not possible to increase the dataset to that size, we had to change plans. Then we worked on the system we mentioned in the methods section, as the first part of the system we needed to train a consistent and reliable VQA model.

We considered two different models for visual question-answering task, the first is the one we coded ourselves, and the second is the state-of-the-art ViLT question answering model. Then, we fine-tuned the ViLT model to our own unique dataset, which we observed that the results of the ViLT model were better. Fine-tuning is a common practice in deep learning where a pre-trained model is tuned on a specific task on a specific dataset. This process includes training the model with a lower learning rate.

Table 2. Evaluation metrics and results of our VQA model

Step	Training Loss	Validation Loss	Wups	Accuracy	F1
100	5.825	5.196	0.099	0.056	0.000814
200	5.173	4.809	0.166	0.109	0.003068
300	4.785	4.457	0.195	0.135	0.004584
400	4.489	4.266	0.220	0.166	0.006706
500	4.296	4.136	0.233	0.181	0.009696
600	4.131	4.050	0.245	0.194	0.010304
700	4.020	3.985	0.256	0.205	0.011724
800	3.957	3.950	0.251	0.199	0.010981
900	3.849	3.918	0.260	0.208	0.011631
1000	3.828	3.901	0.261	0.209	0.012060

Table 3. Evaluation metrics and results of ViLT VQA

Step	Training Loss	Validation Loss	Wups	Accuracy	F1
100	5.624	5.043	0.128	0.072	0.001654
200	5.104	4.956	0.198	0.134	0.005872
300	4.675	4.565	0.235	0.184	0.009894
400	4.264	4.134	0.275	0.218	0.013882
500	4.072	4.102	0.305	0.245	0.017200
600	3.774	3.884	0.325	0.272	0.020462
700	3.632	3.652	0.338	0.296	0.022872
800	3.428	3.456	0.347	0.314	0.024812
900	3.322	3.385	0.352	0.324	0.026542
1000	3.210	3.216	0.356	0.344	0.028325

Table 4. Evaluation metrics and results of ViLT VQA fine-tuned.

Step	Training Loss	Validation Loss	Wups	Accuracy	F1
100	6.532	6.321	0.135	0.094	0.001972
200	5.722	5.697	0.212	0.168	0.007502
300	5.212	5.118	0.268	0.236	0.013528
400	4.482	4.526	0.305	0.292	0.019072
500	3.928	3.972	0.366	0.344	0.024021
600	3.572	3.668	0.382	0.382	0.028012
700	3.234	3.382	0.392	0.414	0.031524
800	2.952	3.106	0.398	0.433	0.034204
900	2.758	2.798	0.402	0.452	0.036523
1000	2.682	2.698	0.416	0.482	0.039052

Training loss is a measure of the error or “loss” of the model on the training set. It measures how far the predictions are from the ground truth values in the training set. Lower training loss means better model performance on the training set.

Validation loss measures the error or “loss” of the model validation set, which is a separated data that not used in training process. Lower validation loss means better model performance on unseen data.

Wups (Wu-Palmer Similarity) measures the semantic similarity between predicted and ground truth answer. It is based on how similar the concepts of the answers in a linguistic sense. Higher Wups score indicates more linguistic similarity between predicted and actual answers.

Accuracy means the proportion of questions which the model predicted answer exactly same with ground truth answer. An accuracy of 1.0 means perfect model, 0.0 means model failed to predict all answers. Higher accuracy indicates better model performance.

F1 score is a measure of model’s accuracy on a dataset. It is the harmonic mean of precision and recall. Precision is the correctly predicted results divided by number of all positive results, and recall is the number of correct positive results divided by the number of positive results. The F1 score is a balanced measure of both precision and recall. Higher F1 scores indicate better model performance.

We considered 2 different models for visual question answering, the first is the one we coded ourselves, and the second is the state-of-the-art ViLT question answering model. Then, we fine-tune the ViLT model to our own data, which we observed that the results of the ViLT model were better.

Figure 13. Training and Validation Loss over steps

7. CONCLUSION

In conclusion, the development of a visual dialog system intended to help visually impaired individuals in understanding the visual world around them. The task offered challenges and learning opportunities. Our initial approach was to create a unique dataset and merge it with the DAQUAR dataset. After building our dataset we implemented various encoding architectures such as hierarchical recurrent encoder, late fusion encoder, and memory network encoder. Even though we optimize hyperparameters for each model during the training phase, we failed to achieve accurate predictions.

After reviewing our initial approach, results led us to limited size of our dataset, an issue we could not fix due to certain constraints. We needed a new approach to maximize our accuracy with our available dataset.

We adopted a strategy involved using fine-tuned versions of several existing state-of-the-art models ViLT VQA, ViT image-captioning, and YOLO object detection. The output of these models was then fed to the ChatGPT-API. This approach was used designed to generate human-like meaningful dialogs that our initial models were unable to generate.

Our journey showed the importance of adaptability in machine learning projects. The limitations to increasing the size of the dataset forced us to think about a new idea that uses a unique solution that leverages the strengths of multiple models. Although our system is still far from perfect, we have managed to achieve reasonable results using a unique design with limited resources.

REFERENCES

Abraham, S. & Li, L. Z. (2018) Visual Dialog IEEE Tansaction on Pattern Analysis and Machine Intelligence

Das, A., Kottur, S., Gupta K, Singh, A., Yadav, D., Moura, J.M., Batra, D. (2017). Visual Dialogue. ECCV

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Berg, T. L., Parikh, D. (2015). VQA: Visual Question Answering International Conference on Computer Vision (ICCV)

Goyal, G., Parikh, D. R., Berg, T. L. (2016). Exploring Models and Data for Image Question Answering. Proceeding of the IEEE conference on Computer Vision and Pattern Recognition.

Kim, H., Lee, J. Y., Lee, J. K. (2016). Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. IEEE Conference on Computer Vision and Pattern Recognition

Kukleva, A. M., Kuznetsov, L. P., Lomonosov, M. A., Nikonorov, A. V., Nikoronova, T. A., Fedrorov, I. O., Goldgof, D. B. (2017). FusionNet: A deep fully residual convolutional neural network for image segmentation in connectomics. Medical Image Analysis.

Ramakrishnan, N., Zhour, B., Li, Y., Olivia, A., Huang, T. S. (2020). VIT: Vision Transformer. arXiv preprint arXiv:2003.10152.

Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.

Raza, A. A., Jain, A. K., Rai, A. N. (2019). A multimodal approach for visual dialog systems for visually impaired people. IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS).

Raza, S. A., Jain, A. K., Rai, A. N. (2020). Visual Dialog systems for visually impaired people: A survey. IEEE Access.

Williams, J. D., Lopes, N. J. (2021). End-to-End Optimization of Goal-Driven and Visually Grounded Dialogue Systems. arXiv preprint arXiv:2012.13356.

Turan, M. T., Al-Regib, A., Hoffman, M. D. (2020). Deep Learning for Vision-Based Assistive Technology: A Survey. Pattern Recognition.

Bolya, D., Zhou, C., Xiao, F., Lee, Y. J. (2019). YOLACT: Real-time Instance Segmentation. ICCV.

Lu, J., Batra, D., Parikh, D., Lee, S. (2019). ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. arXiv preprint arXiv:1908.02265.

APPENDICES