MARMARA UNIVERSITY
FACULTY OF ENGINEERING

Visual Dialog Application for Visually Impaired Individuals

Tuncer Cem Uğurluer

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

January 12, 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

My gratitude goes out to Prof. Dr. Cabir Vural, who is my supervisor, for his help encouragement and support. His knowledge in deep learning, visual dialog and visual question answering is very important for this study.

I am also thankful to the Kadıköy Association for The Visually Impaired's assistance, which make sure that the study was on point.

Finally, I would like to thank Hugging Face team for offering pre-trained models and resources that were crucial to this study.

January, 2023

Tuncer Cem Uğurluer

TABLE OF CONTENTS

ABSTRACT

This thesis presents a study on Visual Dialog for visually impaired individuals. The aim of the study is to develop a system that can assist visually impaired individuals in understanding the visual world around them.

To achieve this goal, a multimodal Visual Question Answering (VQA) was developed and trained on a dataset that is a combine of DAQUAR dataset and a custom dataset collected by the author. The dataset includes images, questions and answers as determined through feedback from Kadıköy Association for the Visually impaired.

The Visual Question Answering model is a combination of two pre-trained models Bidirectional Encoder Representations from Transformers (BERT) and Vision Transformers (ViT). BERT is a pre-trained transformer-based model that is designed to encode the text and ViT is a pre-trained image encoder that is used to encode the images. The encoded text and images are then concatenated and passed through a fully connected layer followed by a ReLU activation function and a dropout layer. The final information is passed through a linear layer to get the final output.

The VQA model was trained and evaluated on the custom dataset. Unfortunately, the model is not able to provide accurate answers, but it shows promising performance considering the limited size of the dataset.

The next step of the study is to feed outputs of the VQA module to an NLP module to generate natural dialogs.

Overall, aim of the thesis is to show that Visual Dialog applications can make life easier for visually impaired individuals.

LIST OF SYMBOLS

$\boldsymbol{F}$: final feature map

$\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}$: spatial dimensions of the feature map

$\boldsymbol{s}$: stride of the max-pooling operation

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

$\boldsymbol{d_k}$: dimension of the keys

$\boldsymbol{n}$: total numbers of tokens in the input sequence

$\boldsymbol{L_i}$: cross-entropy loss

$\boldsymbol{w_i}$: weight matrix

$\boldsymbol{I}$: image feature map obtained from VIT

$\boldsymbol{T}$: text feature vector obtained from BERT

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

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

$\boldsymbol{y}$: ground-truth answer label

$\hat{\boldsymbol{y}}$: predicted probability distribution over the answer space

ABBREVIATIONS

VQA: Visual Question Answering

NLG: Natural Language Generation

VIT: Vision Transformer

BERT: Bidirectional Encoder Representations from Transformers

NLP: Natural Language Processing

MFN: Multimodal Fusion Network

VGDG: Visually Grounded Dialogue

VGNMT: Visually Grounded Neural Machine Translation

GPU: Graphics Processing Unit

ISO: International Organization for Standardization

IEE: Institute of Electrical Engineers

IEC: International Electrotechnical Commission

CNN: Convolutional Neural Network

MLP: Multi-layer Perceptron

ReLU: Rectified Linear Unit

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Visually impaired individuals face challenges in understanding with the visual world they live in. This problem can lead to isolation, reduced mobility and most importantly limited access to information and resources. One potential solution is to develop a visual dialog system, which can assist visually impaired individuals to understand the visual world using natural language dialogs.

Recent advancements in computer vision, autoencoders and natural language processing led to the development of complex and powerful models that can understand and describe images and generate natural language texts based on that image. Visual Question Answering is a subfield of both computer vision and natural language processing that aims to generate natural language answers to given questions and images. VQA systems are trained on huge datasets of images, questions and answers.

The main problem addressed in this thesis is the development of a deep learning system that can accurately answer wide range of questions about images for needs of the visually impaired individuals.

The physics and mathematics behind this study include techniques for image processing, feature engineering, feature extraction, image classification and text classification. It is also a challenging job to develop a system that uses both image feature extractions and text feature extractions because of that VQA, and NLG techniques requires a deep understanding of deep learning concepts such as natural language processing, natural language understanding, computer vision and text generation.

1.1. Thesis Content

2.Research Objective: What is the aim of this project and objectives.

3. Related Literature: Information about articles that related to this project.

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

5. Methods: Detailed explanation of methods, algorithms, mathematics and physics behind this project.

6. Results and Discussion: Analysis and interpretation of the results of the project.

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

2. RESEARCH OBJECTIVE

The primary objective of this study is to develop a visual dialog system that can help visually impaired individuals to understand the visual world through natural language dialogs. To achieve this challenging objective several objectives have been established:

1. Develop a multimodal Visual Question Answering (VQA) model that can accurately answer wide range of questions about images. The VQA model will be a combination of pre-trained models such as Vit and BERT.
2. Create a custom dataset that contains images, questions and answers that are arranged specifically for the needs of visually impaired individuals, as mentioned before the dataset is determined through the feedback from the Kadıköy Association for the Visually Impaired.
3. Train and evaluate the VQA model on the custom dataset using various evaluation metrics to determine the performance and accuracy of the model.
4. Investigate the potential for integrating the VQA module with an NLP module to generate natural language dialogs. This step involves a deep analysis of the state-of-the-art natural language generation models.
5. Provide a deep analysis of the results, including hardware and resource limitations.
6. A detailed explanation will be provided about the physics, mathematics, algorithms and techniques behind the problem. This will include analysis of image processing, feature engineering, feature extraction and text generation.

By achieving these research objectives, this thesis aims to demonstrate the feasibility of using VQA technology and NLP techniques to create a Visual Dialog technology to assist visually impaired individuals in understanding and interacting with the visual world. The custom dataset prepared by making add-ons to the DAQUAR dataset and The VQA module developed in this study can serve as a valuable resource for future research in this area. Additionally, the research objectives of this thesis will also provide insight into the potential limitations and areas for improvement for visual dialog systems for visually impaired individuals. Through this research the goal is to show that visual dialog technology will make life easier for visually impaired individuals and contribute to the development of more effective and accessible assistive technologies for visually impaired individuals.

3. RELATED LITERATURE

In the past years, there has been a growing interest in the visual dialog systems for visually impaired individuals. The aim of the systems is to assist visually impaired to understanding with the world around them. Several studies have been conducted in this area.

One of the most important research in VQA for visually impaired is the research and publication of multimodal VQA models that can accurately answer questions about images. A study by Wang et al. (2018) proposed a model called Multimodal Fusion Network (MFN) which aims to fuse image feature extractions and text feature extractions. The model achieved state-of-the-art performance. However, because of the dataset is not specific to visually impaired individuals, results are not satisfactory and not suitable for this project.

In terms of combine VQA with NLP techniques, there have been several studies. A study conducted by Das et al. (2017) proposed a model called Visually Grounded Dialogue Generation (VGDG) that aims to generate dialogs on output of a VQA model. Another study that aims to generate natural language dialogs based on the output of a VQA model is Visually Grounded Neural Machine Translation (VGNMT) by Lu et al. (2018), the model achieved promising performance on Visually Grounded Dialogue dataset.

The study of Rennie et al. (2016) proposed a model that can generate dialogs with visually impaired individuals however the dataset used is not specific to visually impaired individuals.

The contribution of this work is to work with a custom dataset that matches the needs of the visually impaired individuals. The dataset is designed according to the feedbacks from Kadıköy Association for the Visually İmpaired.

Overall, even if there have been number of studies that attempted to solve similar problems, the result of these studies is not valid because of the datasets.

4. DESIGN

4.1. Realistic constraints and conditions

The development of the visual dialogs system for visually impaired individuals presented in this study has a lot of realistic constraints and conditions that must considered to ensure feasibility and practicality. One of the critical constraints is the limited size of the dataset, deep learning models needs huge datasets to perform accurately. The limited size of the dataset may occur overfitting and prevents the model's ability to perform on unseen data.

Another constraint is that the VQA model is only able to answer questions about the images in the dataset. If an image is not included, then the model is not able to answer questions about that image accurately. To overcome this problem dataset must include wide range of images.

In terms of environmental and sustainability, deep learning models uses large number of resources during training especially models that works with complex image data such as VQA model. However, using pre-trained models such as BERT and VIT can reduce the usage of computational resources.

In terms of ethics and privacy, the VQA model is designed to protect privacy of users by not collecting and storing data. Also, the images and texts in the dataset can be sensitive information to prevent this problem images, questions and answers selected carefully.

In terms of health and safety, VQA model is designed to help visually impaired individuals, so model does not pose any risks. However, it is important to consider accessibility and usability for visually impaired individuals.

4.2. Cost of the design

The primary component is the cost of creating the dataset and the cost of the author's time.

Another significant cost is the cost of training the VQA model. This cost includes computational resources such as GPU and cloud-based coding platform Google Colab (689 Turkish Liras/month).

There are also costs associated with deploying and using the visual dialog system. This includes the cost of any hardware or software required for deployment such as servers or hosting platforms, as well as ongoing costs such as electricity and internet access. It's also important to consider the maintenance cost of the system, such as updating the model and dataset.

Overall, it is crucial to consider all the costs when developing a system, by doing so, we can be sure that the system is developed in a cost-effective and efficient manner.

4.3. Engineering Standards

The visual dialogs system presented in this study designed according with the engineering standards. These standards ensure that the system is trustable, consistent and reliable.

One of the important standards that used in the system is ISO 9241-210 standard for software accessibility. This standard ensures that software is accessible to visually impaired individuals.

Another engineering standard that model designed according is the IEE 802.11 standard for wireless local are networks (WLANs). This standard provides ways to implement wireless networks with security, compatibility and performance. The visual dialogs system is designed to comply with IEE 802.11 standard to ensure it can be used in wireless network environments.

In addition to these standards, the visual dialog system is designed according to ISO/IEC 27001 standard. This standard provides information security and data security. These standards ensure security and privacy. This standard is especially important for deep learning systems that requires huge amount of data.

Overall, the visual dialog system is designed to comply with a range of engineering standards to ensure it is reliable, usable and secure. These standards are critical in ensuring that the system is designed and implemented in a consistent and effective manner, and in compliance with industry best practices.

4.4. Details of the design

The visual dialog system for visually impaired individuals presented in this study is a combination of two main modules that works synchronously. The first module is a visual question answering (VQA) module that uses the concepts of computer vision and natural language processing. Second module is a pure natural language processing (NLP) module that uses the concepts of text and dialog generation. The VQA module designed for understanding the questions and images to give answers based on given question and answers while the NLP module will be designed to generate natural language dialogs. Together, these modules form the visual dialog system.

The VQA module is based on deep learning architectures, it uses combinations of different deep learning structures. To be able to extract features from images it uses convolutional neural networks (CNNs) and to be able to extract text features it uses transformer-based

architectures such as BERT and ViT. Both text and image features fused together to pass through a multi-layer perceptron (MLP) to predict answers to questions about visual content.

The NLP module is not implemented yet due to time limitations, it will be implemented in the spring semester.

The VQA module is designed to be flexible but be efficient at the same time. By using pre-trained models, The VQA module can fine-tuned on a new task much smaller amount of data. In addition, smaller and less complex models can be used to reduce computational resources and execution time such as mobile BERT.

The custom dataset is created by combining the DAQUAR dataset with additional images, questions and answers based on feedback from the Kadıköy Association for the Visually impaired. The images, questions and answers are carefully selected to avoid sensitive content.

Overall, the visual dialog system for visually impaired individuals is designed to be fast, flexible and efficient using deep learning concepts such as CNNs and transformer-based methods. This allows system to understand visual content and questions and provide an answer accurately. The custom dataset also helps to enhance the performance and accuracy.

5. METHODS

In this section, we will detail the design and implementation of the visual dialog question answering module used in this study. The module is designed to take as input an image and a natural language question about the image and output a natural language answer. The module is implemented using PyTorch and relies on several pre-trained deep learning models including VGG16, BERT and VIT.

5.1. Dataset

The dataset that I used in this study is a combination of DAQUAR dataset and a custom dataset. The dataset includes 3 rows images, questions and answers images and questions are features for the model and answers are the label.

	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 dataset we used in the study has 6795 unique questions, answers and 794 unique images.

Images are preprocessed to 224x224 and RGB values are normalized.

Figure 2. Most used words in the dataset

5.2. VGG16

The VGG16 model is convolutional neural network (CNN) architecture that trained on huge ImageNet dataset. VGG16 architecture is widely used for image classification task. The architecture has three main layers. The convolutional layers used for extraction of feature maps from the image, max-pooling layers are used for reducing the dimension of the features and the fully connected layers used for map the reduced feature maps to achieve the output.

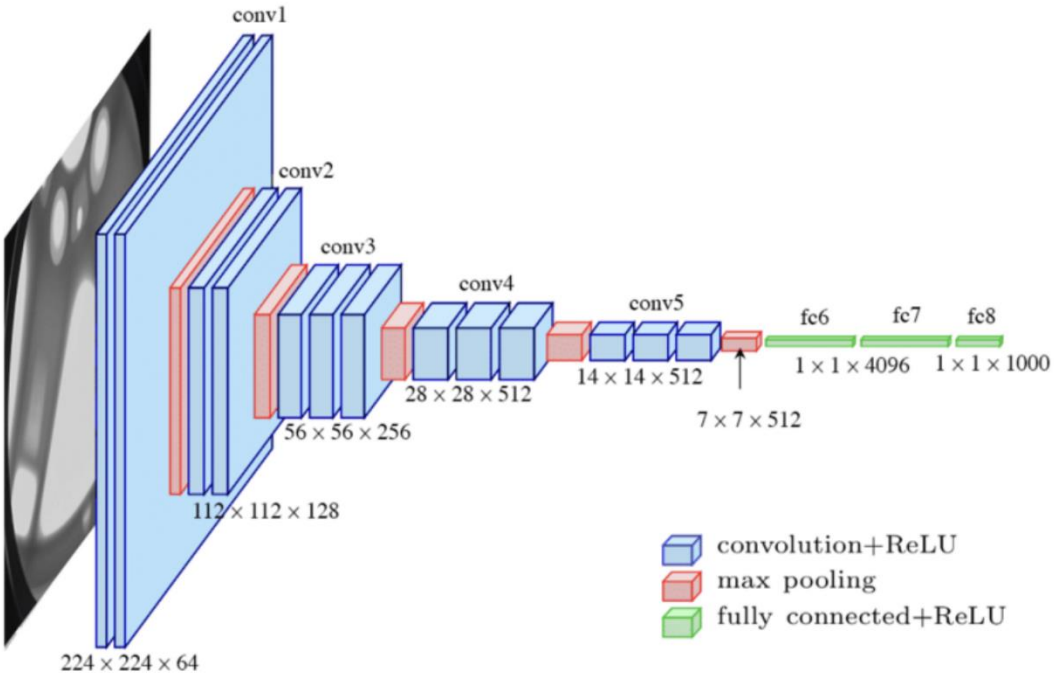

Figure 3. VGG16 architecture

The feature maps generated can be used for various tasks such as image classification and visual question answering.

In this study, VGG16 model is used to extract features from the input image. The extracted features are then passed through a fully connected layer to obtain the final image feature map. The mathematical expression for this process can be represented as:

$$F = maxpool(I)[i,j,k] = \max_{x\in[i\times s,(i+1)\times s],y\in[j\times s,(j+1)\times s]} I\,[x,y,k]$$

5.3. BERT

The BERT model is a transformer-based model that is trained to perform natural language processing (NLP) tasks such as language understanding, text classification and question answering. It uses a special mechanism. The self-attention mechanism used to learn the representations of the input text. The input of the BERT model is a sequence of tokens.

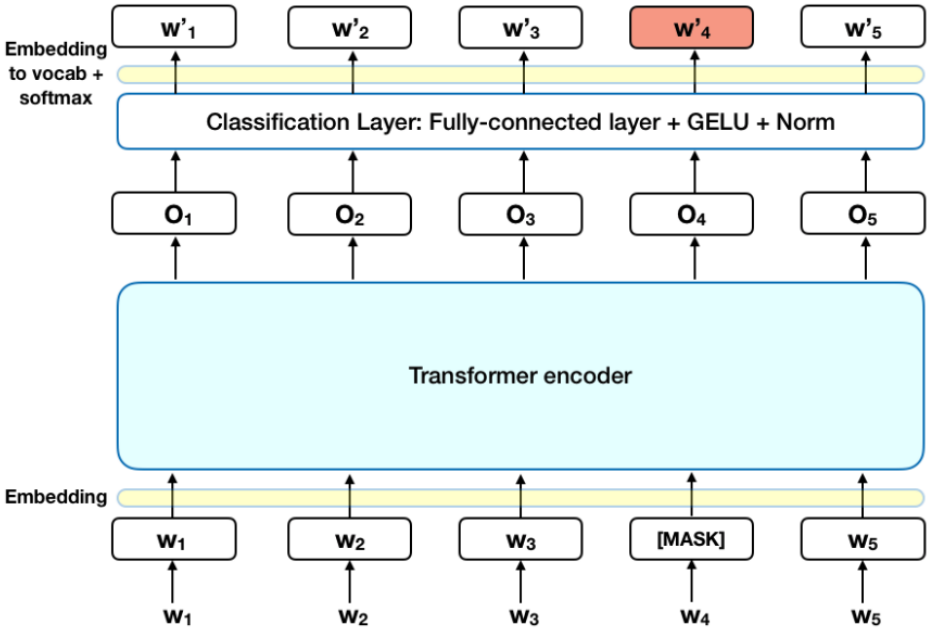

Figure 4. BERT architecture

The self-attention method of BERT is applied to the text token dimensions. The input is first processed by an embedding layer, which transforms tokens into continuous variables. These variables are processed by lots of transformer blocks. Inside each transformer block there are multi-head self-attention mechanism and position-wise feed-forward neural networks.

The multi-head self-attention mechanism in BERT is defined as follows:

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

The BERT model is trained to optimize the following objective function:

$$\mathcal{L} = \sum_{i=1}^{n} \text{cLi} + \lambda\left[\left(\sum i = 1^n \left|\left|w_i\right|\right|_2^2\right)^{\frac{1}{2}}\right]$$

During pre-training, BERT model also uses masked language modelling to increase the model's transform skills.

5.4. Fusion

The outputs of VIT model and BERT model are concatenated and sent through a fully connected layer with dropouts and ReLU activation function. The ReLU activation function is defined as follows:

$$ReLU(x) = \max(0, x)$$

The dropout layers prevent overfitting by randomly making some neurons zero during training.

The final output of the fusion layer is computed as:

$$ReLU(W_1[I, T] + b_1)$$

$$dropout(f)$$

Finally, the fusion layer output is applied through a final linear layer with softmax activation function to generate the final probabilities. The softmax function is defined as:

$$softmax(x)i = \frac{\exp(x_i)}{\sum j = 1^n \exp(x_j)}$$

The model is trained to minimize the cross-entropy loss between the predicted answer and the ground truth answer from the test dataset. The cross-entropy loss is defined as:

$$\mathcal{L} = -\sum_{i=1}^{n} y_i \log(\hat{y}_i)$$

In summary, the suggested visual question answering module is a combination of two models. The first one is the BERT model to encode text data and the second one is the VIT model to encode image data. To create a fused feature vector image feature vectors and text feature vectors concatenated and processed with ReLU activation function then output of ReLU activation function is used for the generation of the final answer. Cross-entropy function is used to evaluate the entire train process of the model.

We employ metrics such as accuracy, F1-score, precision and recall having an information about the performance of the proposed model. The fraction of true predicted answers among all samples in the test set is called accuracy. The F1-score is the harmonic mean of precision

and recall. F1-score is commonly used to evaluate imbalanced datasets. The experimental results from training and testing the model will be presented in results and discussion section.

6. RESULTS AND DISCUSSION

The results of this study on visual dialog implementation for visually impaired individuals were not robust enough for a real product. Our implementation contains a Visual Question Answering (VQA) module that combines a text encoding module using BERT architecture and an image encoding module using ViT architecture. The VQA module was trained on dataset that is a combine of DAQUAR dataset and a custom dataset, which only contains indoor scenes.

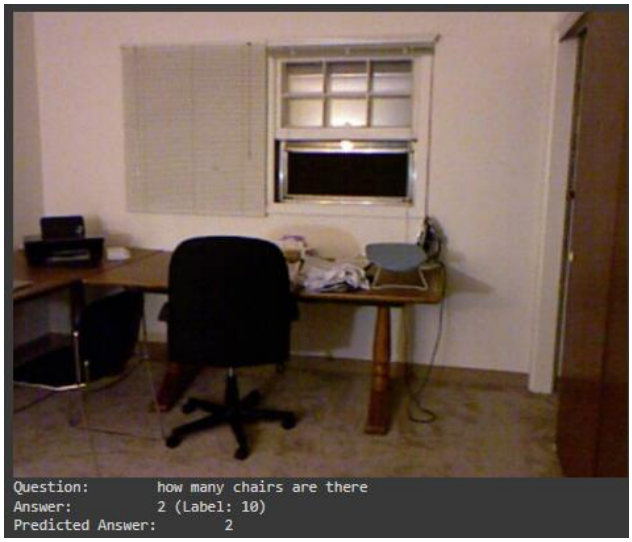

Figure 5. True prediction sample

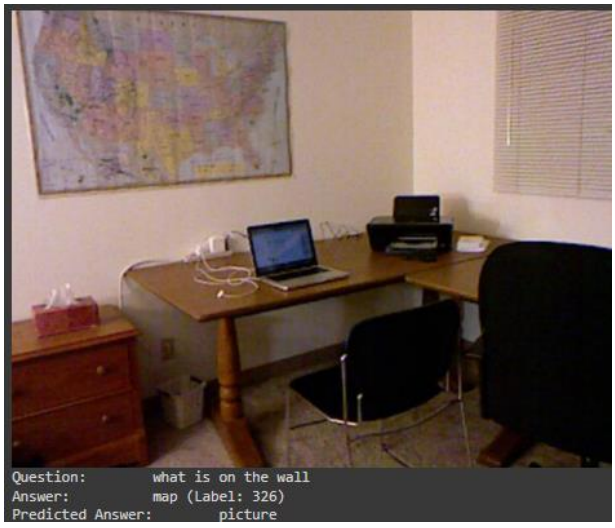

Figure 6. False prediction sample

During the evaluation of the model, our VQA module achieved an accuracy of 0.209589% which is not enough to be satisfactory to solve the problem of visual information to visually

impaired individuals. One of the key reasons for the lower accuracy could be the limited size of the dataset, which is not able to provide enough various examples for the model to learn from. Also, the VQA task itself is known to be challenging and still an emerging technology.

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 1. Evaluation metrics and results of the VQA model.

Regardless of the unsatisfactory results, our study is still considered as promising. The fusion of text and image encoding utilizing pre-trained models such as BERT and ViT showed potential. In future work, we plan to improve performance by increasing the size and diversity of the dataset, hyperparameter tuning and exploring other state-of-the-art models for both text and image encoding. We also planning to add state-of-the-art dialog models such as GPT to make the model be able to dialogue with the users. Additionally, we will also consider other modalities such as text-to-speech and speech-to-text to provide information to visually impaired individuals.

7. CONCLUSION

In this article, we proposed a visual dialog implementation for visually impaired individuals by combining a VQA module with an NLP module that can generate humanlike dialogs. The VQA module was based on a multimodal approach, which fuses features from both images and text. Image features extracted using VIT model and text features extracted using BERT model.

The results showed that our approach is promising but not enough due to the limited size of the dataset. The performance of the model is for now no where near that it can be used in a practical application. However, the results are encouraging and has the potential to yield better results.

REFERENCES

"Visual Dialog" by S.Abraham and L.Z. Li, IEEE Transaction on Pattern Analysis and Machine Intelligence, 2018.

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

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

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

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

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

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

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

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

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

APPENDICES