# A STUDY OF VISUAL QUESTION ANSWERING FOR BLIND PEOPLE

LE THANH TUNG

Japan Advanced Institute of Science and Technology

Doctoral Dissertation

# A STUDY OF VISUAL QUESTION ANSWERING FOR BLIND PEOPLE

LE THANH TUNG

Supervisor : Professor NGUYEN Le Minh

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
Information Science
December, 2021

# Abstract

Multi-media website which contains tons of image and text data has a high demand for extracting and understanding representation and relationship of image and question simultaneously to support users for retrieving information, answering questions, and so on. Besides, it is essential to support blind people as well as the visually impaired community to overcome difficulties in their daily lives. The vision-language systems are promising to learn and understand the visual and textual representation together without the physical vision. Together with its potential, this task also raises some challenges due to unique characteristics of multi-modal systems as well as a specific domain for blind people including i) question may not be in well-grammar texts; ii) image is poor quality from the collecting process that requires a robust approach to extract visual features; iii) unanswerable sample appears the question-answering task.

This study aims to take advantage of advanced Deep Learning techniques to understand and extract meaning and relationship between image and question to predict answers. To this end, the research question is how to employ deep learning architectures to represent and combine the image and question effectively to obtain their hidden relationship especially in the special challenges in VQA dataset for the blind.

To answer the above research question, we propose a hierarchal VQA system including four sub-tasks as follows:

- Answerability Prediction - determines whether the content of images is answered by a question or not, which is useful to eliminate error samples in VQA systems. By taking advantage of Transformer architecture, we propose a VT-Transformer model to extract the visual and textual features delicately thanks to the strength of pre-trained models. According to the experimental results, VT-Transformer generally outperforms the existing baselines. Besides, we also achieve the significant result in VizWiz-VQA 2020 and 2021 competitions.

- Visual Question Classification - divide VQA samples into the specific kinds of questions. Dealing with the difficulties on object-less images, we thus propose an Object-less Visual Question Classification model, OL–LXMERT, to generate virtual objects replacing the dependence of Object Detection in previous Vision-Language systems. Through our experiments in our modified VizWiz-VQC 2020 dataset of blind people, our Object-less LXMERT achieves promising results in the brand-new multi-modal task in comparison to competitive approaches.

- Yes/No Visual Question Answering - solves the specific kind of question instead of all kinds of questions. In this task, we point out the importance of Yes/No question types and propose the BERT-RG model which combines the strength of ResNet and VGG to extract the residual and global features to obtain the visual information. By integrating the stacked attention, the relationship of question and images are intensified by the regional features. Through the detail of experiment and ablation studies, our model outperforms the competitive approaches in VizWiz-VQA 2020 dataset and competition.

- General Visual Question Answering - determines the answer in all kinds of questions. In this work, we propose the novel Bi-direction Co-Attention Network to intensify the textual and visual features simultaneously. Besides, we also apply the VT-Transformer to extract meaningful image and text information. Our method Bi-direction Co-Attention VT-Transformer consistently shows strong performance in the VizWiz-VQA dataset. Besides, it also achieves a promising result in the latest competition in VizWiz-VQA 2021.

Besides the success of each sub-task in the above, our hierarchial VQA system also proves the promising performance against the independent VQA architectures in previous works, especially in VQA for blind people.

**Keywords:** Visual Question Answering, BERT, Vision Transformer, Co-Attention, Answerability, Yes/No Question, VizWiz-VQA, Blind People.

# Acknowledgments

First of all, I wish to express my best sincerest gratitude to my principal advisor, Professor Nguyen Le Minh of Japan Advanced Institute of Science and Technology (JAIST), for his constant encouragement, support, and kind guidance during my Ph.D. course. He has gently inspired me in researching as well as patiently taught me to be strong and self-confident in my study. Without his consistent support, I could not finish the work in this dissertation

I would like to thank Professor Satoshi Tojo, Associate Professor Kiyoaki Shirai, Associate Professor Shinobu Hasegawa of JAIST, and Associate Professor Truyen Tran of Deakin University for useful discussions and comments on this dissertation. Besides, I wish to express my thanks to Dr. Nguyen Tien Huy from the University Of Science, VNU-HCMC for his suggestion and recommendations to enhance our works in the internship.

I am deeply indebted to JAIST for granting me a Doctoral Research Fellow scholarship during the period of my research. I also pass on my thanks to the "JAIST Research Grant for Students" for providing me with their travel grants which supported me to attend and present my work at international conferences.

I would like to thank the JAIST staff for creating a wonderful environment for both research and life. I would love to devote my sincere thanks and appreciation to all members of Nguyen's laboratory. Being a member of Nguyen's lab and JAIST is a wonderful time in my research life.

Finally, I would like to express my sincere gratitude to my parents, sisters, and sweetheart for supporting me with great patience and love. Without their support, I might never complete this work.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Introduction

Due to the fast growth of data over the Internet, especially multi-media information, multi-modals have received a lot of attention in recent years. Most studies focus on understanding and combining many kinds of data such as image and sound, sound and text, and so on. In many kinds of multi-modal systems, Visual Question Answering is obviously promising and practical as a cutting edge task between image and text. Obviously, the amount and number of these modalities including images and texts are extremely massive in the era of multi-media websites. The interest and necessity of this task in both practice and research have encouraged us to put more effort into this challenging area. Among many different vision-text tasks such as Image Captioning, Visual Commonsense Reasoning, the interaction between the human and automatic system in question answering is also the most interesting and fascinating in both research and application.

Especially, Visual Question Answering is completely essential and useful to support blind people with our advanced technologies against their difficulties in their daily lives. Obviously, VQA system is able to understand and predict the answer automatically in the lack of physic vision. Despite the attraction and significance of this topic, there are a lot of challenges against the general VQA domain such as poor-qualified images, complex questions, and unanswerability. The reason comes from the collecting process done by


- **Question:**
  What's the name of this product?
- **Gold Answers:**
  {basil leaves (7), basil (3)}

Figure 1.1: Visual Question Answering Example

blind people. It leads to redundancy and noise in the VizWiz-VQA sample. The detail of these challenges is presented later in Chapter 2. By solving the challenges in VizWiz-VQA, we also raise public attention to difficulties of disabled people, especially blind people. Although our work is not really remarkable among a few pieces of research, we would like to emphasize the need and importance of research to support our community.

Traditionally, VQA system is focusing on the general types of questions which limits the expansion of VQA on the practical side. General Visual Question Answering is extremely challenging and adventurous to overcome as well as deploy it in application. One specific example of VQA sample is presented in Figure 1.1. There are two expected answers for this sample including *basil leaves* and *basil*. Although both of them are correct, the diversity and closeness in the potential answers brings out the difficulties in automatic processing. Depending on the characteristic of VQA task for blind people, we propose a hierarchical framework dividing general VQA systems into many components.

In particular, our systems focus on addressing three main problems in VQA for blind people. Firstly, the poor-qualified samples are eliminated in our first component called Answerability Prediction with the threshold. Then, based on the types of questions, the samples are divided into many specific groups, which is useful to narrow down the potential answer list. With each group, we propose a reasonable architecture based on its own characteristics. Our hierarchical framework allows us to decrease the complexity of VQA systems and enhance each component both independently and dependently.

Technically, VQA tasks are formulated as a classification problem and trained success-

fully via supervised learning methods. However, most of them are based on the Convolution and Recurrent Neural Network which can not take advantage of the recent massive data through fine-tuning. The success of Transformer architecture in recent approaches inspires us to integrate it into our system with the strength of pre-trained models. For each task, we propose a typical architecture capturing its unique specialties to overcome special challenges. In general, the goal of our proposed systems is to digest question and image effectively and extract their relationship to predict the correct answers.

## 1.2 Research Objective and Contribution

The objective of this research is to obtain an effective method for understanding, digesting, and combining the visual and textual features in multi-modal systems, especially in Visual Question Answering. Besides, we also increase the community concern on helping blind people by advanced techniques through our works in the typical dataset. To achieve this aim, the research question is as follows: how to employ deep learning architectures to represent the image and question effective as well as how to combine them to obtain their hidden relationship in VQA dataset for the blind. The emergence of deep learning models has provided an efficient way to learn continuous representation vectors for text as well as image. These representations have a huge contribution to the success of deep learning in many tasks such as Image Captioning [Yun et al., 2019, Sharma et al., 2018], Visual Question Answering [Su et al., 2020a, Tan and Bansal, 2019b, Hudson and Manning, 2019], Visual Commonsense Reasoning [Wang et al., 2020c, Zellers et al., 2019]. Especially, the strength of pre-trained models encourages us to propose and deploy compact and effective approaches to deal with obstacles in the VizWiz-VQA dataset. The research question is answered through four subtasks, which are shown in Figure 1.2, as follows:

**Answerability Prediction** determines the answerability of VQA samples. In this task, with the input of question and image, the Answerability system needs to answer whether this sample is able to answer or not. With our observation in dataset and measurement methods, we propose the novel problem statement in this task. Instead of

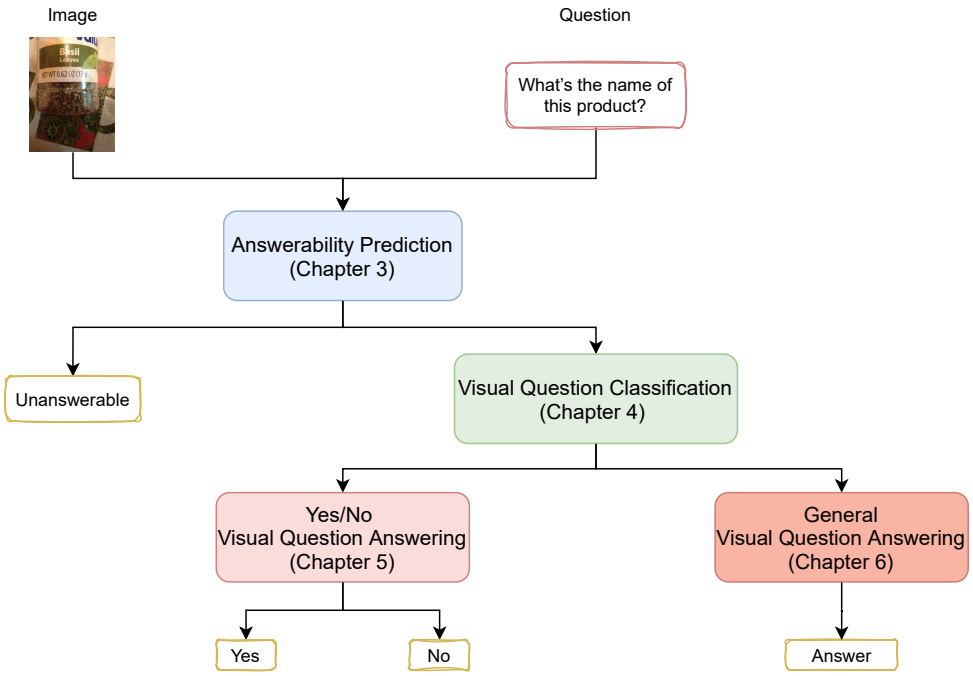



Figure 1.2: The general outline of our hierarchical VQA framework

making a decision about the answerability, our task becomes the regression task to predict the answerability score of VQA sample. The decision of answerability is sent toward the users and their conditions. In this subtask, we proposed a VT-Transformer model to integrate the Vision and Text Transformer delicately. The strength of pre-trained models is obtained effectively in our model to obtains impressive performance in research and application. According to the experimental results, VT-Transformer generally outperforms the approach using CNNs and LSTMs network - the dominant on image and text understanding. Besides, we also achieve the best result in VizWiz-VQA 2020 and 2021 competition in this subtask.

**Visual Question Classification** divides visual questions into their specific categories. With the goal of question's type-driven, Visual Question Classification is advantageous to centralize our following components into the consistent and effective VQA frameworks. Through our proposal in this task, we also make good use of forgotten knowledge in most VQA datasets. Together with the enthusiasm and novelty of Visual Question Classification, we concentrate on is to deal with the weakness of Object Detection on object-less images. We thus propose an Object-less Visual Question Classification model, OL–

LXMERT, to generate virtual objects replacing the dependence of Object Detection in previous Vision-Language systems. Our architecture is effective and powerful enough to digest local and global features of images in understanding the relationship between multiple modalities. Through our experiments in our modified VizWiz-VQC 2020 dataset of blind people, our Object-less LXMERT achieves promising results in the brand-new multimodal task. Furthermore, the detailed ablation studies show the strength and potential of our model in comparison to competitive approaches.

**Yes/No Visual Question Answering** puts the concentration on the specific kind of question. Instead of trying to find the un-limited system, this task proves the importance of Yes/No question types in the Visual Question Answering system. Therefore, we propose the novel problem statements where the system only considers the Yes/No question samples instead of all kinds of questions. In the same research question, we also integrate advanced techniques in Deep Learning to understand image and question together. Traditionally, in this task, Long Short Term Memory (LSTM) and Convolutional Neural Network (CNN) are dominant in question and image embedding. However, the explosion of vanishing gradient in LSTMs and CNNs as increasing the depth encourages us to propose the model which combines the strength of ResNet and VGG to extract the residual and global features to obtain the visual information. With the stacked attention, our model learns the relationship between questions and images to predict the answer for the Yes/No question. According to the experiments, our method achieves competitive results on the real-world dataset for the blind called VizWiz-VQA 2020. The ablation studies and analysis prove the novelty as well as the need for a Yes/No VQA system.

**General Visual Question Answering** determines the answer in the general configuration of VQA. The goal of this task is to predict the answer from the vocabulary for all kinds of questions. This is the general task for most VQA datasets and systems. In this work, we propose the novel Bi-direction Co-Attention Network to intensify the textual and visual features simultaneously. Besides, we also apply the VT-Transformer to extract meaningful image and text information. The consistency in the feature extractor and attention brings us the reduction of computation cost in the attention layer. Our method

Bi-direction Co-Attention VT-Transformer consistently shows strong performances in the VizWiz-VQA dataset. Besides, it also achieves a promising result in the latest competition in VizWiz-VQA 2021.

## 1.3 Dissertation Outline

The remainders of this thesis are organized as follows:

Chapter 2 introduces the problem formulation of Visual Question Answering and Predicting Answerability. We do a detailed analysis of our main dataset and task. Besides, we also present examples to point out the challenges in our tasks. The visualization of samples and label distribution reveals not only the difficulty but also the motivation to encourage us to study this work.

Chapter 3 presents our novel problem formulation of Answerability on Visual Question Answering. Our proposed VT-Transformer takes advantage of two pre-trained models to extract the visual and textual features. We also present the detail of our architecture through the image embedding with Vision Transformer and question embedding with BERT. The experiments are done and evaluated in detail in the real-world dataset for the blind. Through the comparison against the competitive models, our proposed approach outperforms the baselines in both research and competition. We also discuss the results and analysis some ablation studies for making a conclusion about the effect of our proposed architecture.

Chapter 4 mentions our novel multi-modal task on Visual Question Answering for blind people called Visual Question Classification. By utilizing the valuable information in most VQA datasets, we propose OL-LXMERT model to classify a visual question into its category. We also present limits and challenges in most vision-language models with Object Detection as well as our solution. Through our observation, we propose a virtual object generator which exploits visual features to create effective objects in the object-less domain of VQA for blind people. In this part, we also show the detailed comparison and ablation studies of our model against two kinds of competitive baselines, especially in

VizWiz-VQC dataset for blind people.

Chapter 5 describes the novel consideration in Visual Question Answering to focus on the specific kind of question. We also do a literature review to analyze the gaps in current methods. Through our survey and observation, we propose BERT-RG model integrating residual and global information in images into visual features. With the stacked attention mechanism, we combine the regional image features and textual representation into more meaningful space. We also discuss the results and analysis some typical error cases for making a conclusion.

Chapter 6 proposes the bi-direction Co-Attention Network to learn and intensify the visual and textual features simultaneously. We also do a literature review to analyze the gaps in the previous approach. With the bi-directional consideration in Co-Attention Layer, our proposed model obtains competitive results against the existing baselines. Even that, it also achieves promising results in real-world applications. We also present the detail of experiment and ablation studies to prove the novelty of our attention as well as the contribution of our model's components in its success.

Chapter 7 concludes our research and discusses future directions based on our works.

# Chapter 2

# Background

## 2.1 Problem Statement

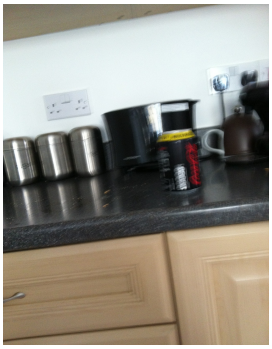
In the multi-media era, it is easy for users to generate their content, which leads to a rapid increase in the amount and type of data. The variety of information makes researchers difficult in exploiting and independently analyzing images, texts, and sounds. The requirements of multi-modal systems which simultaneously solve multi-type of data encourage public interest in new domains. One of them is Visual Question Answering (VQA), which is a cutting-edge topic on image and text.

Visual Question Answering puts the concentration on providing a natural language answer from an image and linguistic question [Shi et al., 2018]. This topic requires an understanding of visual and textual contents via cross-modal models. It is a fundamental study and has attracted many researchers in recent years [Yang et al., 2016, Gurari et al., 2018, Tan and Bansal, 2019b, Kolling et al., 2020, Le et al., 2020].

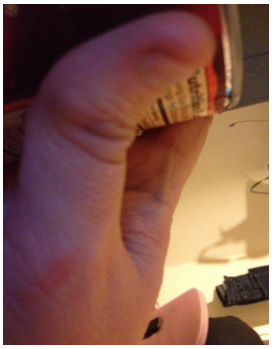
### 2.1.1 Answerability on Visual Question Answering

At the start of VQA for blind people, understanding VQA samples is the paramount objective. Besides the VQA task, predicting the answerability of the VQA sample is also necessary to be concerned about digesting the image and question together. In particular, Answerability on VQA (AVQA) is the task to determine whether a question

can be answered through the content of the image. The examples of AVQA are shown in Figure 2.1.



(a) $VizWiz\_train\_00000001$
**Q:** Can you tell me what is in this can please?
**A:** {soda (1), coca cola (3), coke 0 (2), unsuitable (2), coca cola 0 (1), coke (1)}
**Type:** Other
**Answerability:** 1



(b) $VizWiz\_train\_00000011$
**Q:** What is the sodium content of this can of food?
**A:** {unanswerable (5), unsuitable (4), insufficient photo quality (1)}
**Type:** Unanswerable
**Answerability:** 0

Figure 2.1: The examples of Answerability/Unanswerability in VizWiz-VQA 2020

The reason for unanswerability's existence comes from the quality of images and questions extracted in real-world conversation. In other VQA datasets, a question is generated by annotators from the pre-defined information in images such as objects, colors, and so on. However, in VQA for blind people, images and questions are collected on manual by a personal mobile phone. It causes poor resolution and non-centralization in images as well as the redundant information in questions. This is but also fascinating not only difficult to address AVQA in the VizWiz-VQA dataset. The typical example of an unanswerable sample is presented in Figure 2.1-(b). The geometric center of the image is rotated incorrectly. Besides, the finger of blind people obscures the main object in the image. This phenomenon does not occur in the general VQA dataset. However, it is quite popular in VQA for blind people due to the typical data collection process.

In the VizWiz-VQA challenge in 2020, there is only one team in the AVQA task. It indicates that AVQA is a challenging and brand-new task in VQA. It is, however, a pretty strong motivation for us to study on. Through our observation in the VizWiz-VQA

dataset, we propose to consider AVQA as a regression problem. Indeed, the answerability score is better than an absolute decision. The answerability should be based on the specific domains and the acceptance rate of users.

Mathematically, after extracting the textual and visual features in Equation 2.3, the answerability score is calculated by a regression layer instead of a sigmoid function as follows:

$$AVQA(I, Q) = W_a^T f_{IQ} + b_a \tag{2.1}$$

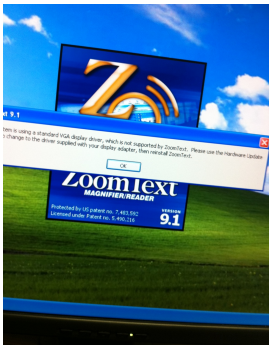
### 2.1.2 Visual Question Classification

In most VQA datasets, it is easy to notice that question type is annotated as a fundamental element. However, this information does not receive any concern in previous works. Instead of accidentally omitting them, we propose a brand-new multi-modal task called Visual Question Classification. The goal of this problem is to determine a category of visual questions. This step can support us to limit the type and number of answers in the practical application.

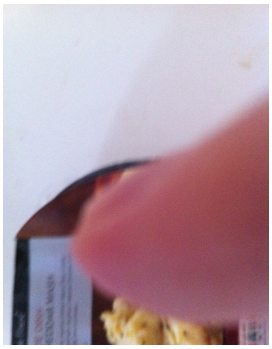
Visual Question Classification (VQC) is quite similar to text classification. However, it requires digesting both images and texts together. This process is more challenging and thrilling than single modal tasks. The typical examples of the VQC task are visualized in Figure 2.2. Obviously, VQC also faces similar challenges to VQA for blind people. Coming from the long-standing existence and enthusiasm in the multi-modal tasks, VQC is worthy to be considered as an independent problem against the traditional text classification.

In particular, VQC systems obtain a pair of an image and a question and classify it into its pre-defined questions' type. Similar to previous works, we also consider VQC as a classification task. In particular, VQC systems need to find a mapping from visual and textual features into the question's type space in Equation 2.2.

$$VQC(I, Q) = Softmax\left(W_{vqc}^T f_{IQ} + b_{vqc}\right) \tag{2.2}$$

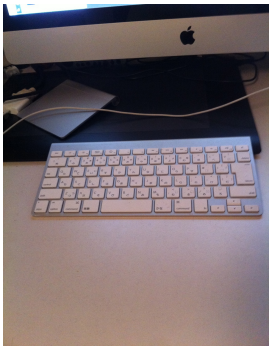



(a) *VizWiz_val_*00000004
**Q:** What does the arrow say?
**QST-Type:** Unanswerable



(b) *VizWiz_val_*00000016
**QST:** What is this?
**QST-Type:** Other

(c) *VizWiz_val_*00003927
**Q:** What does this display say?
**QST-Type:** Number



(d) *VizWiz_val_*00003899
**Q:** Is this a cable?
**QST-Type:** Other

Figure 2.2: The examples of Visual Question Classification in VizWiz-VQA 2020 dataset

### 2.1.3 Visual Question Answering

From its earliest days, Visual Question Answering is formed as a classification problem. The input of the VQA system includes a pair of images and questions, which requires an answer as the output. The answer is in the content of the image such as color, quantity, object, and so on.

Specifically, an example of the Visual Question Answering sample is shown in Figure 2.3. The product image needs digesting to predict the answer as *basil leaves*. It is a great challenge for researchers to design and deploy an automatic VQA system that can answer a difficult question like this.

Mathematically, the Visual Question Answering system is to find a mapping function from the input space of the image and question into the answer space. In most traditional


- **Question:**
  What's the name of this product?
- **Gold Answers:**
  {basil leaves (7), basil (3)}

Figure 2.3: Visual Question Answering Example

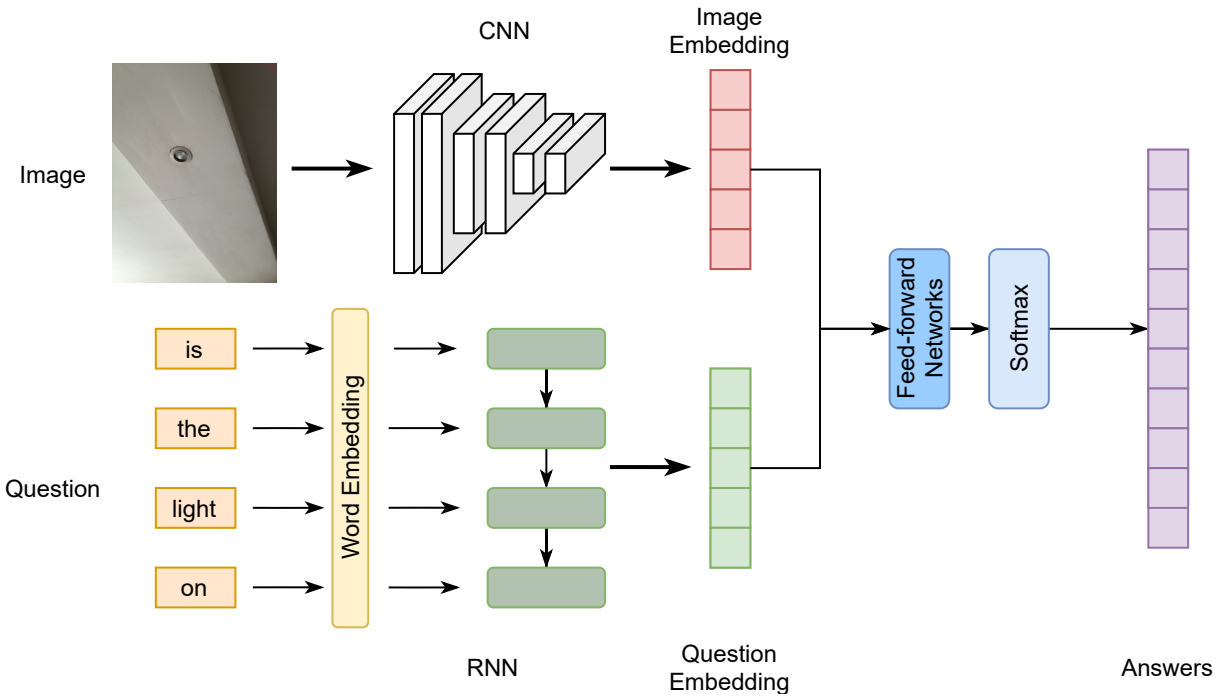



Figure 2.4: General Architecture in Visual Question Answering

VQA system in Figure 2.4, the image $I$ are extracted by a Convolution Neural Networks while the question $Q$ are embedded by a Recurrent Neural Network as follows:

$$f_I = CNN(I); f_Q = RNN(Q) \tag{2.3}$$

Next, the visual and textual features are combined by a vector operation like multiplication, addition, and so on.

$$f_{IQ} = f_I \oplus f_Q \tag{2.4}$$

These composite are used to predict a suitable answer through a classifier whose activation function often is $Softmax$ in Equation 2.5.

$$\hat{y} = Softmax\left(W_a^T f_{IQ} + b_a\right) \tag{2.5}$$

In spite of ten answers in annotation, our VQA system is addressed in multi-class classification. The number of answer vocabulary is presented in each section.

## 2.2 VizWiz dataset

VizWiz-VQA dataset [Gurari et al., 2018, 2019] is published in 2018 and updated until now. Despite an existence of many VQA datasets such as VQA v2.0 [Goyal et al., 2017], CLEVER [Johnson et al., 2017a], VizWiz-VQA is regarded as the first dataset which is aimed to natural conversation [Gurari et al., 2018]. Therefore, we conduct all experiments in VizWiz-VQA 2020 - the latest version of this dataset. From the beginning, VizWiz-VQA originates from an artificial intelligence challenge to help blind people answer visual questions in their daily life. Therefore, it is also meaningful and humane to study this dataset to help our community, especially disabled people.

Specifically, the VizWiz-VQA dataset consists of around 31,000 samples taken by blind people. They take an image by their mobile phone and record their spoken questions. After that, annotators answer the conventional samples as their community contributions. As a result of the daily need of blind people, the quality of images and questions is not really good enough for answering. Even that, there are a lot of unanswerable samples in this dataset. These things get really interesting to study on the real-world data where it exists practical situations. Besides, it means a lot to help our disabled community, especially the blind. By proposing and deploying automatic systems on VizWiz-VQA, we also raise the public interest in both research and industry on the real-life needs of the blind.

Based on the formation of data collection, there are two main tasks including Visual

Question Answering and Answerability Prediction. Other than solving a general VQA problem, VizWiz-VQA also mentions the answerability of these samples. Generally, both all draw towards analyzing, combining and understanding images and questions. The fundamental analyses of the VizWiz-VQA dataset are presented in the following parts. This information is extracted from the VizWiz-VQA Challenge 2019 and 2020. In the framework of this challenge, the evaluation is done online, so it does not provide the test data. Therefore, all our analyses are considered on Train and Validation samples.

### 2.2.1 Answerability on Visual Question Answering

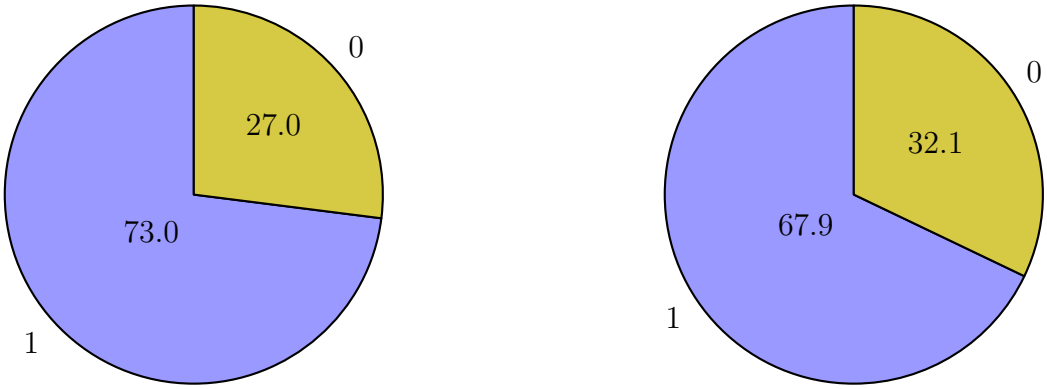
In the first version of the VizWiz-VQA dataset, Answerability on Visual Question Answering is mentioned as the brand-new task to understanding the image and text. This task reflects the important part of VQA samples that are forgotten for many years. The unanswerability of the VQA sample is able to come from the quality of the image, the ambiguity of the question, and even the conflict among annotators. However, in most research VQA datasets, the real-world challenges are almost eliminated by the detailed instruction of annotators in collecting and generating images, questions and answers.

Table 2.1: The detail of Answerability on VizWiz-VQA 2020 dataset

| | **Train** | **Validation** | **Test** |
|---|---|---|---|
| **Answerability** | 14,991 | 2934 | 8,000 |
| **Unanswerability** | 5,532 | 1385 | |

The detail of answerability and unanswerability samples in the VizWiz-VQA 2020 dataset is presented in Table 2.1. Obviously, although the goal of the VizWiz-VQA dataset is about Visual Question Answering tasks, the number of unanswerable is quite large. For the detail of its distribution, we also present the percentage of answerability and unanswerability sample in Figure 2.5.

A large amount of unanswerable samples points out the importance of this task. Instead of forgetting it by considering unanswerable as the answer, the quality of image and question is absolutely worthy to be considered as the independent problem via Answerability on Visual Question Answering.




(a) Train (b) Validation

Figure 2.5: Class Distribution in Answerability on VQA (VizWiz-VQA 2020)

### 2.2.2 Visual Question Classification

Despite popular appearances of question types in many Visual Question Answering datasets, Visual Question Classification is brand-new and worth concern. There is no specific dataset for this task, especially for the object-less image's domain. Therefore, in our experiments, we derive visual question information from an available VQA dataset.

Table 2.2: The analysis of our dataset – VizWiz-VQC

| | **Train** | **Val** | **Test** |
|---|---|---|---|
| **No. Samples** | 16,418 | 4,105 | 4,319 |
| **No. Question Types** | 4 | 4 | 4 |
| **Avg. Words/Question** | 6.76 | 6.74 | 7.26 |
| **Avg. Objects/Image (thresh = 0.4)** | 2.98 | 3.07 | 2.88 |

Together with the consistency in the previous tasks, We also utilize VizWiz-VQA 2020 to create the Visual Question Classification dataset for blind people. The typical characteristic of VizWiz-VQA is about the data process done by blind people, which leads to poor qualify and low resolution in images. Specifically, the number of objects in VizWiz-VQA is approximately 2.9 per image. This ratio is much lower than the other datasets in the same configuration of Faster R-CNN.

Another problem of most popular VQA datasets, as well as VizWiz-VQA, is the concealment of the test set. Therefore, in our novel task, we regard the validation set and 20 percentage of the VizWiz-VQA dataset as the test set and development set in VizWiz-

Table 2.3: The distribution of question type in VizWiz-VQC

| | **Train** | **Val** | **Test** |
|---|---|---|---|
| % Other | 66.91 | 66.92 | 62.31 |
| % Yes/No | 4.67 | 4.65 | 4.51 |
| % Unanswerable | 26.95 | 26.97 | 32.07 |
| % Number | 1.47 | 1.46 | 1.11 |

VQC. The detail of our extracted VQC dataset from VizWiz-VQA 2020 is presented in Table 2.2. We also mention the question type distribution of our data in Table 2.3. The challenges of this dataset are in not only object-less images but also unbalanced labels of question type.

### 2.2.3 Visual Question Answering

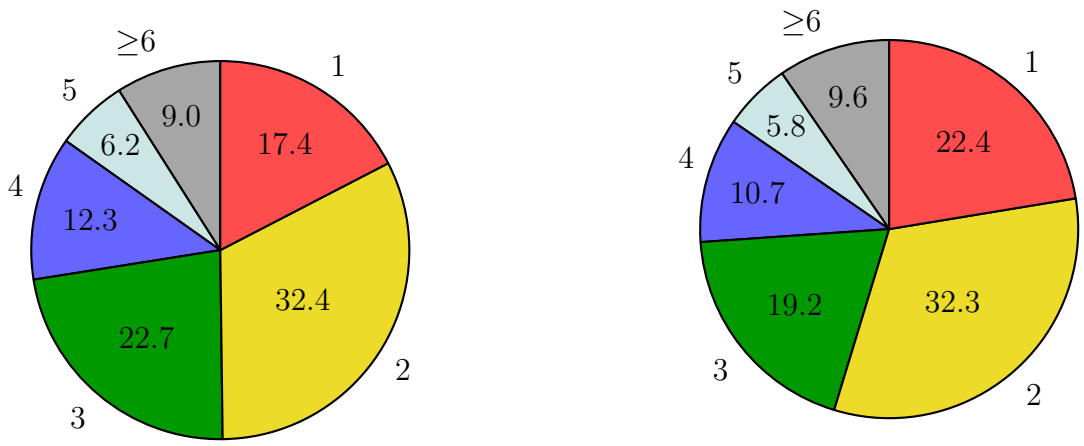
Since its first appearance, the VizWiz-VQA dataset is aimed to solve Visual Question Answering. Similar to the other VQA dataset, each sample includes a pair of an image and a question. For the annotation, there are 10 answers per question. The detail of the question and answer is presented in Table 2.4.

Table 2.4: The detail of VizWiz-VQA 2020 dataset

| | **Train** | **Validation** | **Test** |
|---|---|---|---|
| **No. Image** | 20,523 | 4,319 | 8,000 |
| **No. Question** | 20,523 | 4,319 | 8,000 |
| **No. Answer/ Question** | 10 | 10 | - |
| **No. Answer** | 41,299 | 10,905 | - |
| **Avg. Answer Length** | 3.01 | 2.95 | - |

In most researches, VQA is regarded as a classification task, which exists both advantages and disadvantages. Obviously, the VQA classification system is simpler and easier to deploy than a generation one. This became evident when the length of answers in VizWiz-VQA is around 3 tokens on average.

For more details, we present the distribution of answer length of the VizWiz-VQA dataset in Figure 2.6. In the other sense, the disadvantage of most classification tasks is the out-of-vocabulary (OOV) labels. Assuming all answers in the training set are chosen as the gold labels, there are around 7,500 OOV answers. The error validation samples




(a) Train (b) Validation

Figure 2.6: Answer Length Distribution in VizWiz-VQA 2020

whose labels are not in gold answers is approximately three percent.

Like VQA v2.0 [Goyal et al., 2017], VizWiz-VQA consists of four kinds of question and answer types. There are Yes/No, Number, Unanswerable, and Other. The distribution of question type is shown in Table 2.5. The *Other* and *Unanswerable* samples are quite accounted for VizWiz-VQA.

Table 2.5: The question type distribution in VizWiz-VQA 2020

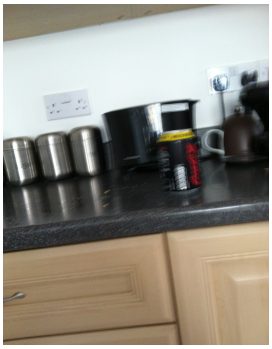
| Question Type | Train | Validation |
|---|---|---|
| **Number** | 301 | 48 |
| **Yes/No** | 957 | 195 |
| **Unanswerable** | 5532 | 1385 |
| **Other** | 13733 | 2691 |

We also present the examples in each question type in Figure 2.7. Despite the distinction of question type, the annotated answers can contain many noise such as *table* in Figure 2.7-(c) of **Number** question.

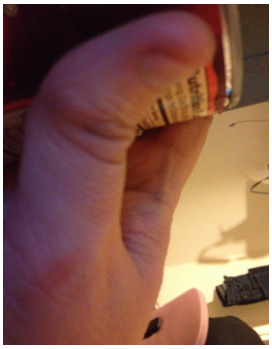
## 2.3 Motivations and Challenges

### 2.3.1 Motivations

In the explosion of multi-media data, the heavy need for the automatic multi-modal system to digest various kinds of information increases dramatically. In this trend, the

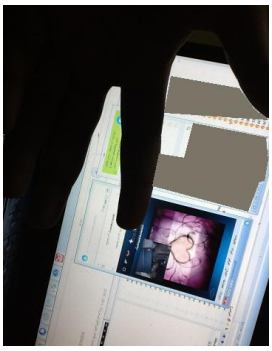


(a) *VizWiz_train_*00000001
**Q:** Can you tell me what is in this can please?
**A:** {soda (1), coca cola (3), coke 0 (2), unsuitable (2), coca cola 0 (1), coke (1)}
**Type:** Other



(b) *VizWiz_train_*00000011
**Q:** What is the sodium content of this can of food?
**A:** {unanswerable (5), unsuitable (4), insufficient photo quality (1)}
**Type:** Unanswerable

(c) *VizWiz_val_*00023845
**Q:** Is this monitor on?
**A:** {yes (9), table (1)}
**Type:** Yes/No



(d) *VizWiz_train_*00023870
**Q:** How many fingers do I have?
**A:** {unanswerable (3), 10, 4 (2), 2 (2), 5, 3}
**Type:** Number

Figure 2.7: The examples of four kinds of question in VizWiz-VQA 2020

combination of image and text is incredible potential. It requires the integration of advantages in both Computer Vision (CV) and Natural Language Processing (NLP), which becomes practical and attractive in research and enterprise. Many tasks in this branch are proposed and deployed such as Image Captioning [Yun et al., 2019, Sharma et al., 2018], Visual Question Answering [Su et al., 2020a, Tan and Bansal, 2019b, Hudson and Manning, 2019], Visual Commonsense Reasoning [Wang et al., 2020c, Zellers et al., 2019], and so on. The main objective of those researches is to digest and combine the information of both image and text.

The practical and scientific interest of this topic motivates us to study the Visual

Question Answering task. By observing many VQA datasets and works, we realize that most recent systems are quite massive and not being used to their potential. In the development of Deep Learning, especially transfer learning, we propose to integrate the strength of pre-trained models into VQA architecture, which makes our system simpler and more effective. The goal of our work is to take advantage of pre-trained resources to increase the understanding of visual and textual features.

Besides the scientific objective, we also aim for humanity purposes in the VizWiz-VQA dataset. Since it was formed in supporting blind people, researches in this dataset make a great contribution to increasing public awareness of disability-friendly applications. There is no denying that VizWiz-VQA is challenging and interesting. The research question in all our works is how to overcome the challenges in VizWiz-VQA to extract, combine, and digest visual and textual features on predicting the correct answers.

### 2.3.2 Challenges

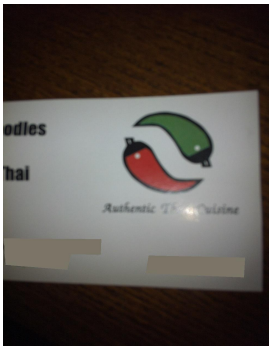
We notice that all samples in this dataset are collected by the blind in their daily activities. It brings us the interest to work in practical and real-world data. However, it also leads to many problems in the VizWiz-VQA dataset. Through our observation, there are two main challenges including (1) Poor Quality Images; (2) Ambiguous Questions. In the rest of this section, examples, and explanations are mentioned to clarify the characteristic of the VizWiz-VQA dataset.

#### Poor Quality Images

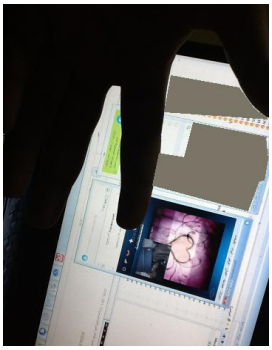
As we mentioned above, all VizWiz images are taken by blind people with their phones. Therefore, the quality of images is not guaranteed. The poor quality image examples are visualized in Figure 2.8. It is easy to realize that these images are collected in the poor condition. The image in Figure 2.8-(a) is even able to recognize nothing while Figure 2.8-(b,c) images are too blurred. In the other sense, Figure 2.8-(d) is a backlit image with the object *"fingers"*. These examples are a great hurdle to extracting the visual features

(a) *VizWiz_train_*00000093
**Q:** What is this?
**Type:** Other

(b) *VizWiz_train_*00000076
**Q:** What kind of soda is that please?
**Type:** Unanswerable



(c) *VizWiz_train_*00023817
**Q:** And does it look like it has water damage done to it?
Thank you.
**Type:** Yes/No



(d) *VizWiz_train_*00023870
**Q:** How many fingers do I have?

**Type:** Number

Figure 2.8: Examples of poor quality images

in VQA. It also prevents us to apply the pre-trained object detection models in our architecture.

### Ambiguous Questions

The second challenge in all VQA tasks is the natural language processing in question. The difficulty and ambiguity in question require more effort to understand. Firstly, there is no constraint to make question sense in the VizWiz-VQA dataset. It means that the object in question is able to appear in the image or not. This characteristic is different from VQA v2.0 [Goyal et al., 2017] dataset.

In the other sense, the questions in VizWiz are recorded by the blind, so it contains more conversational words than the other datasets. Through the comparison in Table 2.6,

Table 2.6: The question length analysis between VizWiz-VQA 2020 and VQA v2.0

| | **VizWiz-VQA** | | **VQA v2.0** | |
|---|---|---|---|---|
| | **Train** | **Validation** | **Train** | **Validation** |
| **Average** | 6.8 | 7.3 | 6.2 | 6.2 |
| **Max** | 62 | 51 | 23 | 23 |
| **Min** | 2 | 2 | 2 | 2 |

the question in VizWiz-VQA is longer than VQA v2.0 [Goyal et al., 2017]. It means that the redundant information in VizWiz-VQA is higher than VQA v2.0. The third question in Table 2.7, for example, are more conversational and practical. It contains a lot of less important and related words such as "Good morning", "Thanks for your assistance". The reason for this phenomenon comes from the data collecting process.

Table 2.7: The examples of ambiguous and redundant questions in VizWiz-VQA 2020

| **Image ID** | **Question** |
|---|---|
| VizWiz_train_00000053 | What is in this card and is it right side up or upside down?<br>Thank you. |
| VizWiz_train_00000063 | what has this picture got in it, what kind of things?<br>What kind of stuff? |
| VizWiz_train_00000075 | Good morning<br>could you please tell me what is in the can in my right hand.<br>Thanks for your assistance |

The second challenge of the question in the VizWiz-VQA dataset is the multiple consideration. In most VQA datasets, one question only refers to a specific object. However, this constraint is not required in the VizWiz-VQA dataset. The first two questions in Table 2.7 includes many different problems. All challenges bring us motivation to study on the VizWiz-VQA dataset instead of the other ones.

# Chapter 3

# Answerability Prediction

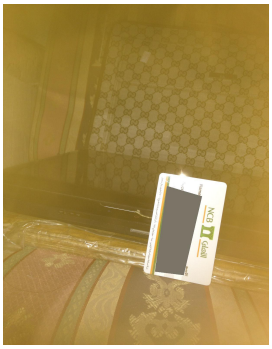
## 3.1 Introduction

In the development of the internet, people need to face tons of multi-media information on every website. It opened a trend of interdisciplinary works in both research and industry. Among them, cutting-edge works in Vision and Language gain more and more interest. The purpose of those studies are to digest and extract the visual and textual representation and their relationship. This viewpoint is easy to meet in many multi-modal tasks such as Visual Question Answering [Su et al., 2020b, Le et al., 2020], Visual Commonsense Reasoning [Wang et al., 2020c], Image Captioning [Yun et al., 2019].

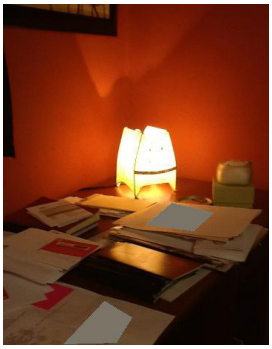
In this work, we put our concentration on an adventurous and brand-new task, Answerability on Visual Question Answering (VQA) appeared recently in a real-world competition of VQA for blind people [Gurari et al., 2018, 2019] in 2018. The reason for unanswerability in VQA samples comes from the quality of images and the content of the question. It is a typical problem in VQA for blind people instead of general VQA tasks. Firstly, we should notice that images and questions in VQA for blind people are taken by themselves. With their vision impairment, it is impossible to control their viewpoint to capture the correct images. Furthermore, questions are generated by their daily needs instead of professional annotators as the previous VQA datasets.

Generally, the goal of this task is to reveal the answerable score that reflects whether

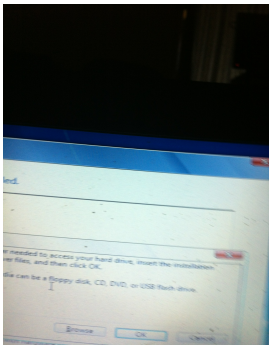
a VQA sample can be answered or not. It is completely new and strange to detect this characteristics. In most VQA datasets, if a sample can be answered, it should be eliminated. However, VQA for blind people is typical and challenging. Together with the answerable sample, there are a lot of unanswerable thing due to the weakness of vision from blind people. It is highly necessary and practical enough to help them with determining this characteristic. If we can quickly respond a VQA sample as an unanswerable one, we can easily answer users to capture it again.



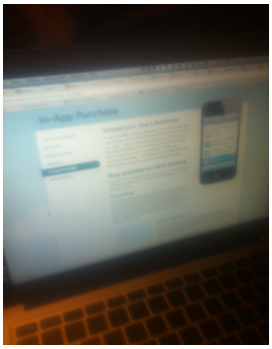
(a) $VizWiz\_train\_00023912$
**Q:** is it not?



(b) $VizWiz\_train\_00023918$
**Q:** What will I do tomorrow?



(c) $VizWiz\_val\_00000058$
**Q:** I cannot move the camera slightly closer to the monitor.



(d) $VizWiz\_train\_00000062$
**Q:** What website is this?

Figure 3.1: The typical examples of unanswerability in VQA for blind people where the gold score is zero.

Traditionally, this task is often based on the aid of VQA systems. The current performance of this task is derived from the outputs of VQA approaches. In particular, in a classification VQA system, the vocabulary is often expanded with the special label, "unanswerable". If a sample belongs to this class via VQA prediction, its answerability score is zero and otherwise. In this consideration, the answerability score of samples is

only one or zero, which is too far from the problem's essence. Coming from our observation on evaluation metrics and the characteristic of VQA for blind people, we propose to regards Answerability Prediction as a regression model instead of a binary mapping from VQA systems. Particularly, this consideration is strengthened by two following reasons as follows:

- Evaluation metrics: The popular measurement in Answerability is the average precision (AP) scores determined by thresholds in a precision-recall curve. This metric is to summarize the precision-recall curve into a single value representing the weighted sum of precisions at each threshold where the weight is the increase in recall. It reflects the balanced point among the pre-defined thresholds in the VizWiz-VQA challenge.

- Optimization goal: The purpose of Visual Question Answering systems and Answerability Prediction is totally different. Answerability reflects conflict of images and questions to predict whether the visual information is meaningful enough to answer or not. Typical examples of unanswerable samples are presented in Figure 3.1. The common problems, for example, are poor resolution in images, ambiguity in linguistic questions, and inconsistency among annotators. Instead of finding the relationship between keywords in question and objects in images, Answerability VQA focuses on the inconsistency in the inputs.

Lately, BERT proves significant successes of Transformer [Vaswani et al., 2017] architectures in Natural Language Processing (NLP). Through the powerful model and huge dataset, BERT is promising to extract the linguistic features in texts. The appearance of BERT in natural language processing is a revolution for text understanding. It also spreads very rapidly into many areas. Accordingly, there are more and more Computer Vision (CV) approaches integrated Transformer architectures and their components [Dosovitskiy et al., 2021, Parmar et al., 2018]. Different from texts, an image is a composite of regional objects instead of a string of words.

Together with the strength of Transformer in CV and NLP, we propose a Vision-Text Transformer combining the successes of Transformer to overcome challenges in Answerability via regression consideration. Besides, our architecture is also integrated by pre-trained models to inherit the performance of original models from the other huge datasets. In experiments, our model outperforms competitive baselines in the VizWiz-2020, a typical dataset for blind people.

Our main contributions of this works are presented as follows:

- We introduce a novel problem statement in Answerability, which is firstly introduced in the research. By considering Answerability on VQA as a regression task, our system reflects the appropriate characteristic of this task.

- We propose a Vision-Text Transformer model to deal with Answerability Prediction task. By taking advantage of the Transformer model, we integrate the pre-trained Vision and Text Transformer into our Answerability VQA model. It is a novel integration in multi-modal tasks to extract visual and textual features simultaneously via Transformer architecture.

- Our proposed model proves its strength and performance in the typical and real-world dataset for blind people, VizWiz-VQA 2020. Through the automatic evaluation tool, our model achieves the state-of-the-art result on this task of the VizWiz-VQA 2020 competition. Besides, our model also outperforms some existing approaches. Through the detailed analysis and ablation studies, our architecture also proves its effectiveness and correctness as integrating Vision and Text Transformers.

## 3.2 Methodology

Traditionally, the goal of all Visual Question Answering is to extract and combine the visual and textual features effectively. However, the challenges in VizWiz-VQA require a powerful model to discover meaningful information from the image and question. In

our architecture, we take advantage of pre-trained models which are optimized in a large dataset to inherit their strength.

Similar to the popular framework in VQA, our model also focuses on two important components including Question Embedding and Image Embedding. In our assumption, we perceive that meaningful features are the key to success in most VQA systems. In our architecture, we propose to integrate the pre-trained language and image classification models into our embeddings. At first glance, although it has less novelty, it is delicate to realize and design it into a complete system. Our originality is to propose an effective regression Answerability on the Visual Question Answering model in the development of Deep Learning.

### 3.2.1 Question embedding

In most linguistic systems, text understanding is always attractive and challenging. In most researches, the biggest question is representing a sequence of texts into meaningful and numeric space. Generally, questions are a sequence of words. Traditionally, word representation is in either a non-contextual or contextual consideration. Non-contextual approaches consider a token as a specific vector in every context. Otherwise, contextual approaches can capture the meaning of words from their context. It allows one word to have multiple linguistic senses.

Furthermore, the explosion of textual data has arisen a need for pre-trained language models. In word representation, there are some popular and famous pre-trained embedding as Word2Vec [Mikolov et al., 2013] and Glove [Pennington et al., 2014]. These approaches, however, have their weaknesses. In particular, they are focused on word representation instead of sentences and documents. Specifically, Word2Vec and Glove are non-contextual word representation which does not consider the context of words in the sequence.

Recently, the appearance of BERT [Devlin et al., 2019] and Transformer architecture [Vaswani et al., 2017] marked a significant moment in linguistic representation. Un-

doubtedly, the strength of BERT comes from its approach based on unsupervised learning. It allows BERT to learn from a huge of raw datasets without the requirement of manual annotation. Context-free models such as Word2vec [Mikolov et al., 2013] and GloVe [Pennington et al., 2014] represent a specific and fixed word representation in every sentence and document, where BERT takes into account its context. BERT allows one word to have multiple linguistic senses. In the real world, the word meaning depends on its context. This characteristic is firstly expressed in BERT. Therefore, there is no doubt about the success of BERT, especially in text understanding.

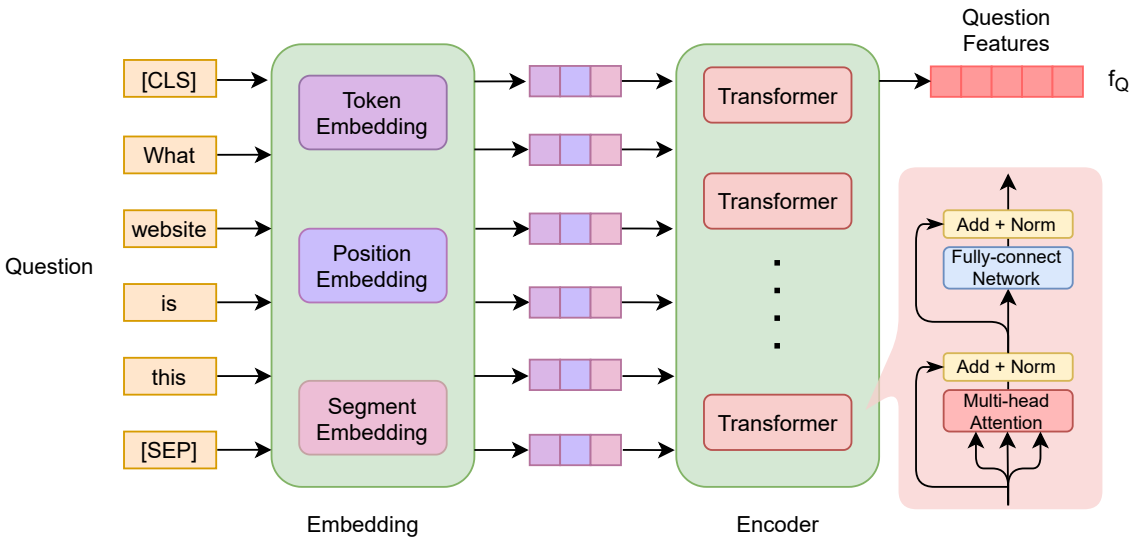



Figure 3.2: Question Embedding: extracts textual features of words in question by BERT - Text Transformer.

Specifically, an input sequence is supplied with a special $[CLS]$ token which represents the starting position of the sentence. In the same way, $[SEP]$ token is inserted in the last position of the sequence to mark the ending position. Similar to vanilla Transformer, BERT gets a modified sequence as input to put into through the stack of Encoder. In each block, the input is applied by the self-attention, layer normalization, and feed-forward network to generate the meaningful features through the gradual evolution of textual signal.

The detail of our embedding is visualized in Figure 3.2. Specifically, after embedding layer, the sentence is represented $s \in \mathbb{R}^{n \times d}$. The Transformer Encoder is the trainable function to map the embedding space into the textual one via the Multi-head attention,

Layer Norm, and Feed-forward layers. In each encoder block, the previous input are augmented through its self-attention signal in Equation 3.1 and Equation 3.2.

$$u_i^{'} = MH - Attention(s_i) \tag{3.1}$$

$$u_i = LayerNorm(s_i + u_i^{'}; \gamma_1, \beta_1) \tag{3.2}$$

where the LayerNorm function is defined in Equation.

$$LayerNorm(z; \gamma, \beta) = \gamma \frac{\left(z - \frac{1}{k}\sum_{i=1}^{k} z_i\right)}{\sqrt{\frac{1}{k}\sum_{i=1}^{k}\left(z_i - \frac{1}{k}\sum_{i=1}^{k} z_i\right)^2}} \tag{3.3}$$

In the next step, the signal is transmitted into the Feed-forward layer and normalized by another LayerNorm layer as follows.

$$z_i^{'} = W_2^T ReLU(W_1^T u_i) \tag{3.4}$$

$$z_i = LayerNorm(u_i + z_i^{'}; \gamma_2, \beta_2) \tag{3.5}$$

In our architecture, Question Embedding takes advantage of the pre-trained BERT model to extract the sentence representation. In this task, our question is processed in BERT in one segmentation. Traditionally, through the flow of BERT, the question representation is derived from an output of $[CLS]$ vector at the last layer of the Transformer.

### 3.2.2 Image embedding

With the rapid rise of images, it is essential to have new technology that is better than Convolution Neural Networks(CNNs) in the traditional approaches. With the development of quantum computing and device, the need for huge computational processing is not really problematic. Currently, most vision tasks utilize the strength of GPUs/ TPUs

and a massive amount of datasets. The recent successes of Transformer in NLP inspire researchers to integrate it into the vision area [Carion et al., 2020, Wang et al., 2020a]. The interest of Vision Transformer is its fewest modification from the original one. Instead of deploying a different architecture, this model uses the vanilla Transformer model with the new kind of input embedding from images.

Inspired by the strength of Transformer in Computer Vision, we propose to integrate the Vision Transformer [Dosovitskiy et al., 2021] into our image embedding. By utilizing the massive datasets, the pre-trained Vision Transformer is ideal to replace the traditional visual feature extractor. In the benchmark dataset in the image as ImageNet, the Vision Transformer has incredible performance against the traditional models. Therefore, instead of the wasted effort in building the powerful architecture, the recent approaches move toward the fine-tuning mechanism where the pre-trained models are used in the fundamental components to deploy the systems in the other tasks effectively and quickly.

Unlike texts, images contain much more information in them basically in form of pixels. The explosion of computation cost of the consideration where the pixel is similar to the token is very huge and unavailable to deploy even with current hardware. Therefore, the image is broken into many fixed-size patches. The patches are then flattened and sent for further linear projection with the position embedding. Specifically, the input image $x \in \mathbb{R}^{H \times W \times C}$ is split into many patches whose resolution is $(P, P)$ as follows.

$$x = \left[x_p^1, x_p^2, ..., x_p^N\right] \tag{3.6}$$

After that, the resolution of the image becomes $x \in \mathbb{R}^{N \times (P^2 \times C)}$ where $H, W, C$ is the height, weight, and channel of $x$. Following the configuration of BERT, the special token $x_{class}$ is also pre-pended into the start of the patches to present the class representation. The series of patches and positions are presented in Equation 3.7.

$$z^0 = [x_{class}; x_p^1 E; x_p^2 E; ...; x_p^N E] + E_{pos} \tag{3.7}$$

where $E \in \mathbb{R}^{(P^2 C) \times d}$ and $E_{pos} = \mathbb{R}^{(N+1) \times d}$

At each Transformer blocks $l$, the representation of patches are calculated by Equation 3.8 and Equation 3.9.

$$u^l = MH - Attention(LayerNorm(z^{l-1})) + z^{l-1} \tag{3.8}$$

$$z^l = \left(W^T(LayerNorm(u^l)) + b\right) + u^l \tag{3.9}$$

where $MH - Attention(q, K, V)$ is calculted by Equation 3.10 and Equation 3.11

$$MH - Attention(q, K, V) = [head_1, head_2, ..., head_h]W^O \tag{3.10}$$

$$head_j = softmax(\frac{qW_j^Q(KW_j^K)^T}{\sqrt{d}})VW_j^V \tag{3.11}$$

The equation of MH-Attention and Layer Norm layer is mentioned in the previous section. After the processing in the stack of Transformer blocks, the image is mapped in the series of $N + 1$ d-dimensional features.

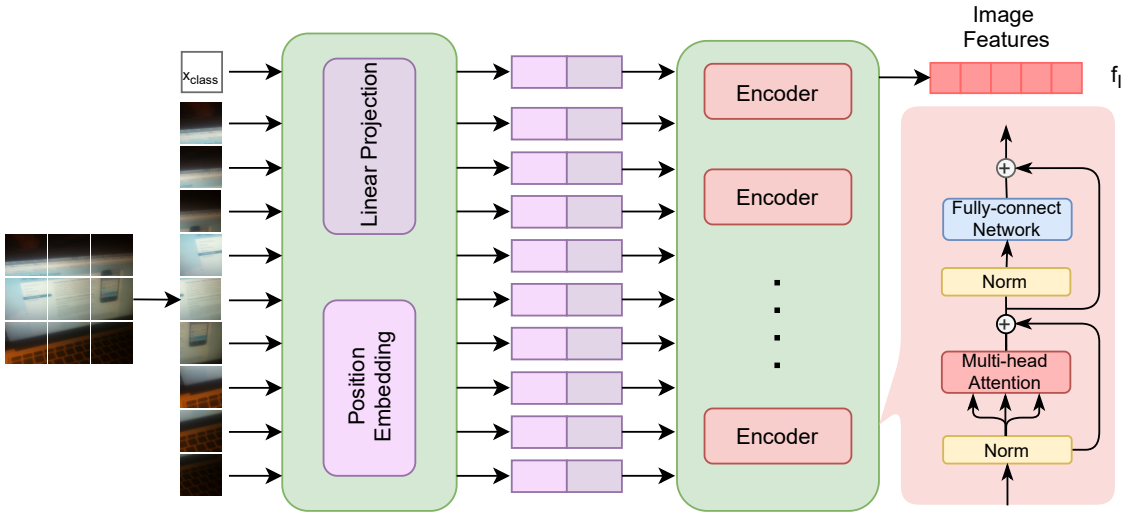



Figure 3.3: Our image embedding integrates Vision Transformer to extract the regional and visual features.

In our architecture, we integrate a pre-trained Vision Transformer into our complete system as the Image Embedding phase. The detail of our image embedding is presented

in Figure 3.3. In particular, we modify the original classifier in Vision Transformer into a fully-connected layer to map the visual features into fixed-dimension space. Through the combination of textual features, we make use of pre-trained knowledge in ImageNet in our task. Instead of learning from scratch, our model is highly leveraged the success of the pre-trained image classification models. Via our embedding, the representation of the image is extracted by the $[x_{class}$ vector at the last layer of Transformer blocks.

### 3.2.3 Vision-Text Transformer Model

After integrating the powerful visual and textual feature extractor, we propose a novel framework to deal with this task - Answerability on Visual Question Answering. As we present above, the goal of this task is to predict the score to reflect whether the VQA sample is able to answer or not. From our assumption, meaningful features play an important role in the success of most VQA systems. Therefore, in our work, we just use the simple multi-modal fusion function to reduce the explosion of trainable parameters in two Transformer components.

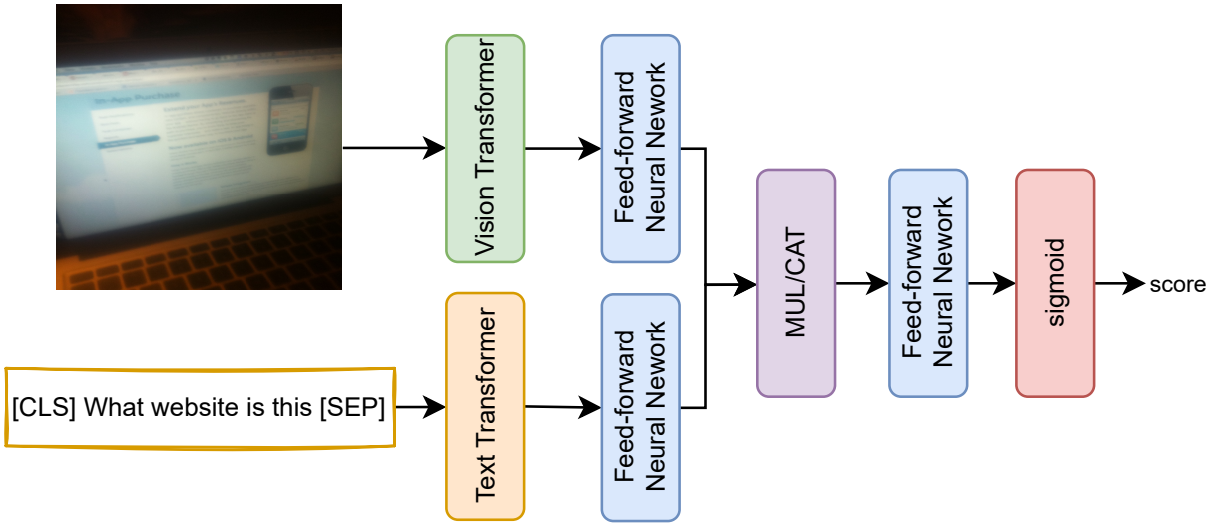



Figure 3.4: The detailed architecture of VT-Transformer for Answerability Prediction

Specifically, in our architecture, image and question features are mapped into a fixed-

dimension space by a fully-connected layer in Equation 3.12.

$$f_I^{'} = W_I^T f_I + b_I; f_Q^{'} = W_Q^T f_Q + b_Q \tag{3.12}$$

where $W_I, W_Q \in \mathbb{R}^{k \times d}$ and $b_I, b_Q \in \mathbb{R}^d$ are trainable parameters. Generally, the image and question are mapped from the original dimension $\mathbb{R}^k$ of Transformer into the consistent $d-$space.

These layers are also integrated into the original feature extractor at the last layer of prediction. Obviously, we also utilize the representation of $[CLS]$ on question embedding and $[x_{class}$ as our meaningful features.

After the huge computation in the feature extractor, we use a multi-modal fusion function to combine the visual and textual features. As we mentioned above, our goal is to make the fusion function as simple as possible. Therefore, we propose to use two operations including either multiplication (MUL) or concatenation (CAT). Accordingly, the visual-textual representation is calculated in either Equation 3.13 or Equation 3.14.

$$f_{IQ} = MUL(f_I, f_Q) = f_I \cdot f_Q \in \mathbb{R}^d \tag{3.13}$$

$$f_{IQ} = CAT(f_I, f_Q) = [f_I || f_Q] \in \mathbb{R}^{d+d} \tag{3.14}$$

Finally, the answerability score is determined in a regression model instead of a classification model. Due to the range of answerability scores, we propose to use the sigmoid function to normalize the prediction into $[0, 1]$. Mathematically, the answerability score is calculated in Equation 3.15. The dimension of matrix $W_r, b_r$ depends on the fusion function.

$$score = \hat{s} = \sigma(W_r^T f_{IQ} + b_r) \tag{3.15}$$

With our novel problem statement as regression model, we use Mean Squared Error (MSE) in Equation 3.16 as our loss function instead of cross-entropy in VQA.

$$Loss = \frac{1}{N}\sum_{i=1}^{n}(s_i - \hat{s_i})^2 \tag{3.16}$$

Where $s_i = \{0, 1\}$ corresponds to unanswerable and answerable sample in annotation. Thanks to a feedback signal from MSE loss, we optimize vision and text Transformer to be as close to the distribution in our local dataset.

## 3.3 Experiment

### 3.3.1 Dataset

VizWiz dataset [Gurari et al., 2018, 2019] appeared in 2018 and annually updated until now. In this dataset, Answerability Prediction is firstly considered as an individual task. Unfortunately, this task is replaced by Image Captioning in the VizWiz-VQA 2020 and 2022[1]. Interestingly, despite less interest in the contest, this task has still existed since the beginning of VizWiz-VQA challenges. Obviously, Answerability on Visual Question Answering also attracts attention despite its challenges and lack of public awareness.

Table 3.1: Detail of VizWiz 2020 dataset in Answerability

| Dataset | Train | Validation | Test |
|---|---|---|---|
| **No. Samples** | 20523 | 4319 | 8000 |
| **No. Answerable (%)** | 14981 (73%) | 2937 (68%) | - |
| **No. Unanswerable (%)** | 5542 (27%) | 1382 (32%) | - |

The specific information of this dataset is presented in Table 3.1. The percentage of unanswerable samples is quite high by 27% in train and 32% in the validation set. It is very popular in practice, especially for blind people. Therefore, it is essential and practical to deal with Answerability Prediction. This task is an important component in VQA systems to eliminate unanswerable samples before answering them.

[1] https://vizwiz.org/workshops/2020-workshop/

### 3.3.2 Evaluation Metric

Based on the configuation of VizWiz-VQA dataset, the test set is not provided. All evaluation should be doned by an online frameworks in EvalAI[2]. Therefore, it has restricted the manual modification in the automatic systems. According to VizWiz-VQA Challenge, **Average Precision** is used as the main evaluation measurement which is calculated by the weighted mean of precision under a precision-recall curve in Equation 3.17.

$$AP = \sum_{n} (R_n - R_{n-1}) \; P_n \tag{3.17}$$

where $R_n$ and $P_n$ are the precision and recall at the $n$-th threshold. The pre-defined thresholds are not revealed to the public.

The second metric we would like to consider is F1-score which measures the balance between precision and recall. However, we do not implement any evaluation metrics here. All experimental results in this work are extracted from the online system, EvalAI. Therefore, these results are really reliable and credible to be considered.

### 3.3.3 Experimental Settings

All details of our implementation are presented in Table 3.2 to reproduce our model. As we easily realize, we use the basic setting of both Vision and Text Transformer. The reason for our choice comes from the explosion of computation cost. Besides, we also prove that this setting is effective and ideal enough for a powerful Answerability Prediction system in practice.

Table 3.2: Detail of experimental settings

| Components | Value |
|---|---|
| **Vision Transformer** | B_16_imagenet1k. |
| **BERT** | bert-base-uncased |
| **Full-connected Layer** | CAT: 768 - 512 - 1024 - 512 - 1<br>MUL: 768 - 512 - 512 - 512 - 1 |
| **Vector operation** | multiplication (MUL)/concatenation (CAT) |
| **Optimizer** | AdamW(lr = 3e-5, eps = 1e-8) |

[2] https://eval.ai/web/challenges/challenge-page/743/overview

### 3.3.4 Results

As we mentioned above, most previous Answerability systems are based on the results of VQA approaches. Therefore, it is too hard to find any existing publication of this task. Therefore, we propose to compare our work with four competitive baselines as follows:

- The best performance in EvalAI Leaderboard 2020 - VWTest[3]: VWTest obtains the best result in the Answerability task from the online system. Without its publication, the result in the competition and public evaluation system is really reliable to be considered as the existing baseline. In this comparison, we also emphasize the importance of real-world results in blind evaluation.

- Classification + Mapping: Traditionally, Answerability is based on the result of general Visual Question Answering. The answerability score is 1 if the answer is not "Unanswerable" and otherwise. Through this consideration, we would like to consider the two most popular and powerful baselines in the General Visual Question Answering task.

  - CNN-LSTM [Goyal et al., 2019, Antol et al., 2015]: This model uses the basic CNNs layers for image embedding and LSTM with Glove for question embedding. Despite its simplicity, it is effective enough to be mentioned in every VQA dataset as the strong baseline. We also use the Pytorch implementation in Github [4] to reproduce the results.

  - FT-VQA [Kazemi and Elqursh, 2017]: In the developments of pre-trained models in Computer Vision, this model takes advantage of the Resnet-152 model to extract the visual features. Similar to the previous approaches, LSTM with Glove embedding is also utilized for question understanding. All our results for this model are based on the re-implementation [5] in Pytorch

- BERT-RG [Le et al., 2020] Regression: This model has the best success in the

[3] https://eval.ai/web/challenges/challenge-page/743/overview
[4] https://github.com/ntusteeian/VQA_CNN-LSTM
[5] https://github.com/Cyanogenoid/pytorch-vqa

Yes/No binary classification. Therefore, it is promising to work well in regression tasks. Besides, the strength of BERT-RG is to combine both residual and global features from ResNet and VGG. Two of them are typical in Convolutional approaches remaining dominant in image understanding. This characteristic is suitable for our work to compare Vision Transformer against traditional methods. Based on the original architecture of BERT-RG, we modify the binary classifier into the scoring function in regression.

Table 3.3: The comparison results in Answerability on VizWiz-VQA 2020 dataset

| Type of Model | Model | Average Precision | F1-score |
|---|---|---|---|
| **No Infor. + Real-world** | **VWTest Leaderboard 2020** | 26.84 | 42.32 |
| **Classification + Mapping** | **CNN-LSTM** Goyal et al. [2019] | 32.14 | 44.47 |
| | **FT-VQA** Kazemi and Elqursh [2017] | 29.85 | 42.91 |
| **Regression** | **BERT-RG Regression** Le et al. [2020] | 52.22 | 41.85 |
| | **VT-Transformer (Our model)** | **76.96** | **67.26** |

Apparently, our consideration in the regression task is more suitable than traditional approaches. All results of the Average Precision in Regression Model are better than the classification ones. It proves that our problem statement is meaningful and important to re-form the new approaches. On the other hand, we can not deny the strength of our proposed model, VT-Transformer. Our model outperforms all baselines in both Average Precision and F1-score. The enhancement of our works is also promising and potential.

### 3.3.5 Ablation Studies

In this part, we also conduct some ablation studies of components in our architecture to prove our novelty in this work. In our comparison, we would like to emphasize the strength of our model against the powerful and popular components in previous approaches. Generally, there are three studies as follows:

- Vision Transformer against CNNs in Image Embedding
- The robustness and effectiveness of pre-trained models
- The trade-off between capacity and accuracy

Firstly, we would like to reveal the strength of Vision Transformer into Image Embedding. In this component, we compare Vision Transformer against two powerful image classification models that consist of ResNet [He et al., 2016] and VGG [Simonyan and Zisserman, 2014]. These image models belong to CNNs which is dominant in Computer Vision for many decades. This comparison not only reflects the strength of the Vision Transformer against the existing methods but also reveals the robustness and effectiveness of the Transformer in the new downstream task.

In experiments, we only consider the pre-trained models instead of training from the scratch. Specifically, we use pre-trained parameters of ResNet-152 and VGG-16 in the Pytorch library [6]. For Vision Transformer, we keep the same configuration in Table 3.2. In this experiment, we also mention the effect of fusion operations consisting of multiplication (MUL) and concatenation (CAT). The detail of our results in this ablation study is presented in Table 3.4.

Table 3.4: Ablation studies on Image Embedding modules

| Fusion Operation | Model | Average Precision | F1-score |
|---|---|---|---|
| **MUL** | **ResNet152** | 63.13 | 63.59 |
| | **VGG16** | 70.75 | 66.82 |
| | **Vision Transformer** | **76.96** | **67.26** |
| **CAT** | **ResNet152** | 67.74 | 65.16 |
| | **VGG16** | 71.57 | 63.62 |
| | **Vision Transformer** | **74.91** | **66.70** |

In this comparison, VT-Transformer outperforms ResNet and VGG by approximately 10% on Average Precision and 3% on F1-score. It proves that Vision Transformer works well on Answerability instead of traditional image models. Transformer architecture brings the success of feature extraction in both vision and text. In somehow, consistent

[6] https://pytorch.org/vision/stable/models.html

architectures in image and question embedding lead to the effectiveness in optimization. On the other hand, although all models are pre-trained in ImageNet dataset [Deng et al., 2009], Transformer architecture is more robust and effective to deploy in a new task, Answerability. Besides, it is easy to realize that multiplication is better than concatenation in this task.

Secondly, we also reveal the effects of pre-trained parameters in deploying our Answerability system. Due to the success of multiplication in the above study, we only consider it as a vector fusion function to combine visual and textual features in this comparison. Firstly, VT-Transformer is inited by pre-trained parameters $B-16$ from Pytorch library [7]. After, this model is fine-tuned through the learning process in the Answerability task. On the other hand, the VT-Transformer with no pre-trained weights is trained from the scratch.

The results of the two models are presented in Table 3.5. Obviously, with the same architecture, the model with pre-trained weights is better than another one. Successes of pre-trained systems with the fine-tuning mechanism come from their strong optimization in huge datasets. Based on the image understanding in the classification of the ImageNet dataset, Vision Transformer is able to discover the effective representation of images.

Table 3.5: Effects of pre-trained parameters in VT-Transformer

| Models | Average Precision | F1-score |
|---|---|---|
| **w/o pretrained parameter** | 58.95 | 57.01 |
| **with pretrained parameter** | **76.96** | **67.26** |

Finally, we also present the impacts of different sizes on the Vision Transformer. From this comparison, we would like to emphasize the trade-off between computation cost and accuracy. Obviously, in every application, the balance of the cost and the effectiveness always obtains the interest. In the range of our works and facilities, we only deploy the model called by $VT-16$. This setting is smaller than $B-16$ at all components. With $B-16$, Transformer blocks are 12 and the number of head in multi-head attention is 12

[7] https://github.com/lukemelas/PyTorch-Pretrained-ViT

while the hidden and latent dim are 768 and 3072. The order of configuration is similar to $VT-16$. Obviously, we also pre-train our model, VT-16, in ImageNet. Therefore, all results are based on the model without pre-trained weights.

Table 3.6: A comparison of Vision Transformer architectures

| Model | AveragePrecision | F1-score |
|---|---|---|
| **B-16**<br>**12 - 768 - 3072 - 12** | **58.95** | **57.01** |
| **VT-16**<br>**3 - 512 - 512 - 8** | 56.38 | 47.87 |

The details of our comparison are shown in Table 3.6. The completed size of the VT-Transformer is approximately 2400MB and 460MB corresponding to B-16 and VT-16. The bigger model is, the more capacity it has. Although B-16 is better than VT-16, a smaller version of VT-16 also gets a high performance against the existing baselines in Table 3.3. It proves that our proposed model with Vision and Text Transformer is promising and effective for the Answerability Prediction task.

## 3.4 Summary

In this work, we firstly introduce a new problem statement in an interesting and brand-new task in Visual Question Answering. Instead of going on a classification task, our proposed regression is more practical and related to the main evaluation measurements in its original challenge. By the scoring function, our system is potential and flexible enough to put it into practice via the different thresholds. Besides our novel consideration, we also propose a Vision-Text Transformer model to overcome the challenges of the Answerability task. Our model inherits the strength of pre-trained models to integrate them into our architecture. In addition, we take advantage of Transformer in both image and text to understand and extract visual and textual features effectively. With the simple multi-modal function, our model still proves its efficiency in Answerability on Visual Question Answering. Through the detailed experiments and ablation studies, our model outperforms the competitive baselines in VizWiz-VQA 2020 dataset. The comprehensive

analysis, our model and its components prove its novelty and efficiency through the previous approaches. Besides the success in accuracy, our model is promising to spread the interest in the VizWiz-VQA to support blind people in their daily life.

# Chapter 4

# Visual Question Classification

## 4.1 Introduction

In the era of multi-media, both volume and variety of data have been increasing rapidly. These varieties of data representation bring us convenience and interest in our daily lives. Nevertheless, we need more effort to understand many different modalities (i.e images, texts, and videos), which requires advanced technologies in cutting-edge areas such as Natural Language Processing, Computer Vision, and Speech Processing. In recent decades, multi-modal approaches are highly interesting in many researchers [Lu et al., 2020], Kiela et al. [2020]. Those works focus on analyzing, understanding, and retrieving the relationship among various modalities [Le et al., 2021a].

Together with the success in text and image processing, more and more researchers have arisen their motivation in vision-language tasks. The reason comes from the associated composition between images and texts. In this trend, there are many cutting-edge tasks such as Visual Question Answering [Le et al., 2020], Visual Commonsense Reasoning [Wang et al., 2020c], Image Captioning [Yun et al., 2019], and so on.

In the necessity of vision-language researches, we propose an interesting and novel task named Visual Question Classification (VQC). The goal of this task is to determine the category of visual questions. This kind of question is a mutual interaction between image and text. Obviously, Visual Question Classification plays an important role in Visual

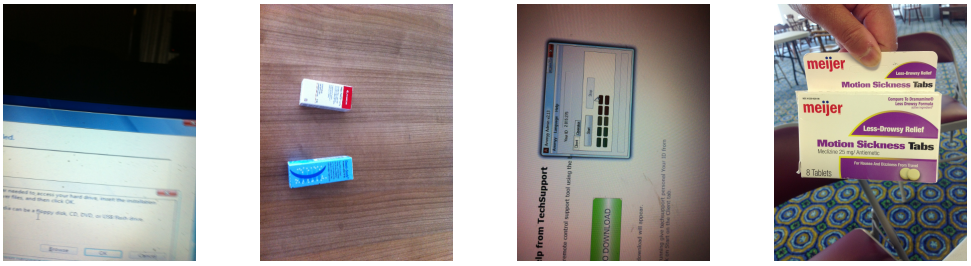



Figure 4.1: The typical examples of low-qualified images

Question Answering systems. With an aid of a Visual Question Classification approach, it is useful to narrow down the searching space of Visual Question Answering systems. Although this is the first time this task is introduced, the category of visual question always exists in most Visual Question Answering datasets such as VQAv2.0 [Goyal et al., 2017], VizWiz-VQA [Gurari et al., 2018], etc. This fact reflects the enormous potential of question type which has not ever been used in the previous approaches.

Recently, the success of transfer learning motivates more and more researchers to deploy the pre-trained models in many areas such as BERT [Devlin et al., 2019] in Natural Language Processing, SpeechBERT [Chuang et al., 2020] in Speech Processing, Vision Transformer [Dosovitskiy et al., 2021] in Computer Vision, etc. These models are fundamental components of many powerful approaches in research and industry. Together with the spread of vision-language researches, vision-language models (e.g. LXMERT [Tan and Bansal, 2019a]) get a lot more attention in recent years. Those models are based on the combination of object-based features via Object Detection models for images and context-based representation via Transformer architecture for texts.

From their earliest days, the pre-trained vision-language models prove their strength in many multi-modal tasks such as Visual Question Answering [Chen et al., 2020], Image Caption [Tan and Bansal, 2019a], Visual Commonsense Reasoning [Su et al., 2020b], Image-Text Retrieval [Chen et al., 2020], etc. However, our concerns come from real-world images in low resolution as well as low-qualified objects. In our work, these kinds of images are regarded as object-less ones. Throughout our work, these images are representative of samples that have few objects recognized from Object Detection models. This kind of

image is more practical and challenging than previous ones in carefully modified datasets such as CLEVR [Johnson et al., 2017b], VQAv2.0 [Goyal et al., 2017], etc. Typical examples of these kinds of images are presented in Figure 4.1. In these cases, it is not effective to utilize Object Detection models such as Faster R-CNN [Ren et al., 2017] to extract the object-based features. Therefore, the important and practical research question is whether the object-based vision-language model is powerful enough to struggle against object-less images.

To overcome the challenges of object-less images, we propose a compact generator to take advantage of Transformer-based image features in constructing the virtual objects. Our object-less generator is pragmatic and effective to replace the role of the Object Detection component in recent vision-language models. Obviously, object-based approaches seem vulnerable due to the resolution of images. Our proposed component is less influenced by the side effects of poor-quality images. It makes use of typical and hidden characteristics in the image's content to create virtual objects. Therefore, our proposed model is easy to adapt into the object-less domain. Together with its generalization, our generator is a promising choice in object-less images. Besides, there is no denying that object-based features are more complicated and effective in the high-qualified environment if Object Detection is adapted into the target dataset. However, the gold resources of Object Detection are more expensive than Image Classification, especially in self-supervised learning. Although our proposed architecture is not limited to poor-qualified vision, it is meaningful to emphasize the strength of our proposed architecture in the specific domain (i.e in the VizWiz-VQA dataset of blind people).

In particular, we utilize Vision Transformer to extract the visual features, which allows us to capture the local and global characteristics of images via image classification tasks. Then, the general image representation is put into our architecture to generate the virtual object's information. Besides, we also integrate our fake objects into LXMERT [Tan and Bansal, 2019a], a well-known vision-language model, to predict the types of visual questions. Through the detailed comparison against existing baselines and ablation studies, our model proves its strength in object-less images and outperforms State-of-the-art

approaches in text classification.

In the consistent motivation of our work, we focus on the VizWiz-VQA dataset for blind people. The typical characteristic of this dataset is its data collection process. All samples of VizWiz-VQA are collected by blind people. Therefore, images in the VizWiz-VQA dataset are often object-less and poor-qualified. Studying in this domain is ideal to raise the public interest for the disabled especially for blind people. It is humane and necessary to help them overcome their difficulties in the real world via deploying advanced technologies, especially in vision-language understanding.

Our main contributions are as follows:

- We introduce a novel task, Visual Question Classification, in the cutting-edge area between vision and language. It is the first time this task is regarded as an independent problem despite its importance and concealed long-term viability.
- Through our observation of images from blind people in the VizWiz-VQA dataset, we indicate that object-less images are formidable challenges in vision-language tasks, especially the object-based approaches. Therefore, we propose our object-less generator to eliminate the massive dependence on Object Detection models.
- We propose the Object-less Visual Question Classification model which takes advantage of image features to generate the virtual objects and integrates the vision-language models to predict the category of visual questions.
- Experimental results and ablation studies on VizWiz-VQA 2020 dataset prove the effectiveness and robustness of our virtual objects against object-based models.

## 4.2 Related Works

### 4.2.1 Text classification

In recent years, with the development of Transformer architecture and transfer-learning, text understanding achieves significant improvements, especially in sentence classifica-

tion [Kim [2014b]]. Well-known models in this trend such as BERT [Devlin et al. [2019]], XLNet [Yang et al. [2019]] is one of the competitive approaches in a lot of text classification datasets. Different from previous autoencoding and autoregressive approaches, XLNet proposes a new framework to learn the context of a word based on the contribution of all tokens via the permutation operation. Generally, the strength of these approaches depends on the self-supervised learning from huge datasets and the robustness of self-attention in Transformer architecture.

In contrast to inductive learning in the above approaches, transductive learning techniques are effective to model the relationship between texts via observing all data samples. In the traditional approaches, the category of text is considered by the local context the global relationship through textual graphs. In this kind of approach, BERT-GCN is the powerful approach that combines a large-scale pre-training language model and transductive learning. Through a heterogeneous graph of textual elements, TextGCN [Yao et al. [2019]] is able to learn the text representation via the relationship matrix of nodes and weighted edges. To integrate external language model, BERT-GCN [Lin et al. [2021]] proposes an ensemble classifier with the contribution of both BERT [Devlin et al. [2019]] and TextGCN [Yao et al. [2019]].

### 4.2.2 Vision-Language Model

In the rapid growth of multi-modal information, vision-language models have been receiving interest from both research and industry. Obviously, the visual content is useful to emphasize the context of documents while textual information is effective to intensify objects and attributes in images. However, the difference of representation between images and texts is a huge hurdle in previous works. In recent years, there is a lot of effort to overcome this challenge in many vision-language tasks such as Visual Question Answering [Le et al. [2020]], Visual Commonsense Reasoning [Wang et al. [2020c]], Image Captioning [Yun et al. [2019]], etc. The general process of these approaches is to learn the visual and textual features independently via the strength of NLP and CV methods and

combine them with the multi-modal fusion function. These systems, however, accidentally ignore the composition between textual and visual objects.

The typical example of this branch is Vision-Text Transformer [Le et al. [2021b]]. This model utilizes Vision Transformer [Dosovitskiy et al. [2021]] to embed an image and BERT [Devlin et al. [2019]] to learn textual features. Then, it combines them via either multiplication or concatenation in vector space. Based on the robustness of Vision Transformer and BERT, this model achieves significant results without massive effort into the multi-modal fusion function. The advantage of this work is based on its simplification and ease of deployment. However, the interaction of texts and images in the feature extraction module is little regarded.

Instead of learning images and texts independently, the pre-training vision-language models are optimized to digest and extract their incorporation simultaneously. One of the most successful vision-language models, LXMERT [Tan and Bansal [2019a]], utilize bi-directional cross-encoder attention to learn the visual and textual features together. In particular, LXMERT utilizes the Faster R-CNN to extract the object-based features of images and Transformer architecture to digest the multi-modal relationship via the huge aggregated VQA datasets. Therefore, the LXMERT model has the power to learn the generalization and composition from two different kinds of inputs. However, its weakness comes from the pre-training dataset where images are guaranteed about their quality and object's quantity. Therefore, the performance of this kind of vision-language model is based on the strength of the Object Detection model and the target environment.

## 4.3 Methodology

The performance of most current vision-language models depends on the quality of the Object Detection system, which is the bottleneck in the practical domain, especially in poor-qualified images. With our observation and the trend of image processing, we propose an object-less generator that utilizes the visual features to eliminate the requirements of the external Object Detection models. Instead of starting from scratch, we take ad-

vantage of pre-trained models via transfer-learning to construct powerful virtual objects. To prove the efficiency of our proposed component, we integrate it into one of the most successful vision-language models, LXMERT Tan and Bansal [2019a].

### 4.3.1 Object-Less Generation

#### Image Embedding

Similar to the previous part, we also take advantage of Vision Transformer with a few modifications as our Image Embedding. The detail of the Vision Transformer in this phase is presented in Figure 4.2.

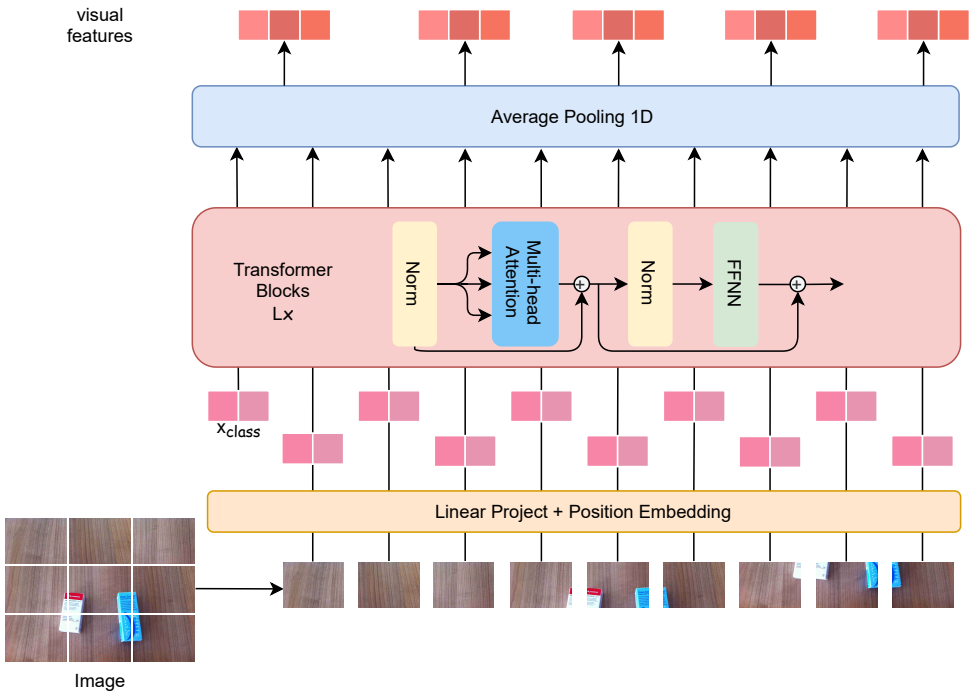



Figure 4.2: The detailed architecture of our Image Feature Extraction

Different from Image Embedding in Chapter 3.2.2, our component utilize all outputs of Vision Transformer as the visual features in Equation 4.1.

$$\{f_{i_k}\}_{k=1}^{K} = ViT\left(\left[x^0_{class}, x^0_1, x^0_2, ..., x^0_N\right]\right) \tag{4.1}$$

Besides, to reduce the computation cost, we also integrate the stack of Average Pooling 1D in the visual features to obtain the generalized representation.

We also notice that Vision Transformer is used as the external feature extractor instead of an internal component in our model, which is similar to most previous image processing approaches. This mechanism is ideal enough to increase the speed of our system.

### Object-less Generation

In our model, we assume that image features are sufficient to generalize visual contents such as objects, colors, backgrounds, and so on. Therefore, the global representation of images is a great alternative to output from an external Object Detection model. Furthermore, annotated data of image classification is easier and more reasonable than object-based resources. Even that, our proposed component is completely accomplished to integrate the portable image system into understanding visual content.

In the previous components, an image is represented into the feature vector $f_I \in \mathbb{R}^{d \times k}$ where d is the dimension of the output layer in Vision Transformer and $k$ is the number of patches in images. In most vision-language approaches, an object consists of two main characteristics extracted by a specific region of interest (RoI). Firstly, with each selected RoI, the RoI features are derived from the mean-pooled convolution layer of the Object Detection model. Secondly, the position of bounding boxes in images is also utilized to represent objects. However, we also consider some thresholds to eliminate the uncertain objects. In traditional approaches, object-based features are often extracted by the process in the Up-Down model [Anderson et al. [2018b]]. Besides, to maintain consistency in an input layer, the number of objects is often equal to 36.

Based on the configuration of previous approaches in image processing, our virtual objects also includes two kinds of features. With each patch $f_i \in f_I$, RoI features of corresponding virtual object $o_i$ are generated by a Feed-forward Neural Network Equation 4.2.

$$r_i = \tanh\left(W_r^T f_i + b_r\right) \tag{4.2}$$

Where $W_r \in \mathbb{R}^{d \times 2048}, b_r \in \mathbb{R}^{2048}$.

The dimension of RoI features is similar to the Up-Down model [Anderson et al. [2018b]] for the original objects. The number of virtual objects is, however, based on the number of patches in Vision Transformer instead of the Up-Down model [Anderson et al. [2018b]].

After generating the RoI features of virtual objects, we also learn the corresponding position via Feed-forward Neural Network models without an activation function in Equation 4.3.

$$p_i = W_p^T f_i + b_p \tag{4.3}$$

Where $W_p \in \mathbb{R}^{d \times 4}, b_p \in \mathbb{R}^4$. In most Object Detection models, we need to determine the specific format of bounding boxes such as width-height, point-length, etc. However, our Position Generator is efficient enough to learn spatial information in all configurations. Similar to RoI Generation, the number of position features are based on the number of patches from Vision Transformer [Dosovitskiy et al. [2021]]. The visualization of our process in RoI Feature and Position Generator is presented in Figure 4.3.

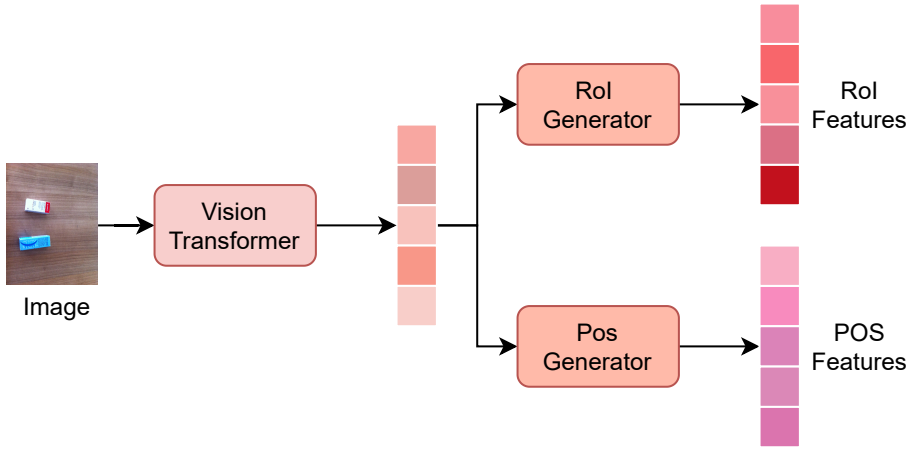



Figure 4.3: The visualization of RoI and Position Generator

At a quick glance, our generation is too easy to consider as the critical component in the novel vision-language models. However, we need to notice that the features of objects depend on the bounding boxes. Therefore, each object is represented locally in each specific region of an image instead of considering the global features. In object-less images, there are a few objects with high confidence. It is the reason that object-based features

dwell on redundant regions in images for low confident objects. Obviously, localization is indeed challenging in CNN-based approaches. In our generation, object features are created by the global representation of images, so our virtual objects are ideal to capture global information and learn local regions in images. Although our generation is simple and transparent, its generalization is quite high and efficient. The performance of our proposed generator is proved in the detailed results and ablation studies.

### 4.3.2 Object-less Visual Question Classification

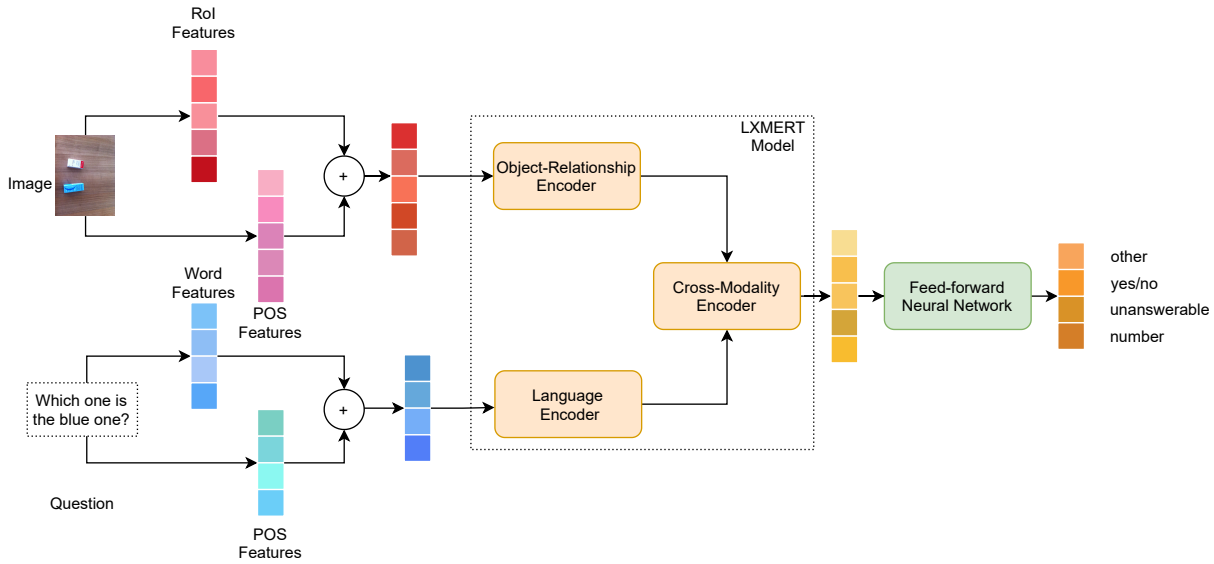



Figure 4.4: OL-LXMERT: The integration of object-less generator and vision-language model for Visual Question Classification

The key component in previous vision-language approaches is to expand the input and embedding layer for processing both image and text simultaneously. However, most of them often depends on the external Object Detection models. With the fewest modification of the vision-language model, we propose a novel Object-less Visual Question Classification model. Our model takes advantage of the pre-trained LXMERT model with our virtual objects. with our proposed architecture, it is ideal to evolve gradually through the development of the Computer Vision and Vision-Language model.

Firtly, after generating the virtual object $o = \{o_i\}_{i=1}^{m}$ from our object-less generator, virtual object features $r_i$ and position $p_i$ are normalized by LayerNorm function (LN).

Then, the position-aware embedding is calculated by adding the information of normalized $r_i$ and $p_i$ in Equation 4.4. From this aggregation, image representation becomes the combination of virtual object features and position information.

$$v_i = \frac{(LN(W_F r_i + b_F) + LN(W_P p_i + b_P))}{2} \tag{4.4}$$

Next, the visual features $v$ and textual features $q$ are intensified by Multi-head attention in Transformer architecture [Vaswani et al. [2017]]. To combine the multi-modal information, LXMERT proposes a cross-modality encoder between images and texts. In particular, this component is the bi-direction multi-head attention from images to texts and vice versa. It is so similar to guided-attention [Yu et al. [2019]] in previous works. However, the success of the LXMERT model comes from the pre-training strategies. This process allows vision-language models to learn the multi-modal data instead of single modality in traditional approaches.

In our architecture, we take advantage of the pre-trained LXMERT model to prove the strength of our virtual objects. Therefore, LXMERT architecture is almost maintained. Instead of utilizing the object-based feature, our OL-LXMERT model obtains the virtual objects via the internal object-less generator. It is pragmatic and effective to deploy in both practice and research. Even that, our proposed generator may completely replace the role of the Object Detection module in vision-language models in this task.

## 4.4 Experiments

### 4.4.1 Datasets and Evaluation Metrics

Coming from the novelty of this task, we recommend utilizing VizWiz-VQA 2020 dataset to create VizWiz-VQC. The detail of our extraction and VizWiz-VQC dataset is presented in Chapter 2.2.2. In evaluation, similar to previous approaches in Question Classification, we also consider F1-score as the main metric in Visual Question Classification. In addition, we also mention the precision and recall of our classifier in all comparisons.

### 4.4.2 Results

Table 4.1: The detailed comparison of our Object-less approach against the competitive baselines

| | Model | Precision | Recall | F1 |
|---|---|---|---|---|
| **Q** | **BERTGCN** [Lin et al. [2021]] | 0.59 | 0.61 | 0.59 |
| | **XLNet** [Yang et al. [2019]] | 0.57 | 0.58 | 0.57 |
| **QI** | **VT-Transformer**[ Le et al. [2021b]] | 0.59 | 0.65 | 0.61 |
| **QI** | **LXMERT** [Tan and Bansal [2019a]] | 0.63 | **0.70** | 0.66 |
| **QI** | **OL–LXMERT (Our model)** | **0.67** | 0.69 | **0.68** |

Coming from the brand-new appearance of VQC in the multi-modal tasks, it is too hard to choose the competitive baselines in this problem. As a pioneer in the task, we suggest comparing the VQA task to two kinds of models including (i) general text classification and (ii) multi-modal VQA. In the first comparison, we mention two SOTA text classification models which are based on the latest technologies of XLNET [Yang et al. [2019]] for pre-trained language understanding and BERT-GCN [Lin et al. [2021]] for Graph Neural Network. These approaches only receive textual questions (Q) as an input instead of both images (I) and texts (Q). In the aspect of multi-modal systems, we compare our model to VT-Transformer [Le et al. [2021b]] which is one of the most successful models in the VizWiz-VQA task. The reason for this choice comes from the similar image feature extractor between the two models. With a few modifications in the last prediction layer, we consider VT-Transformer as the most related approach in comparison to our model. Besides, to prove the strength of our model in the object-less domain, we also present the performance of the original LXMERT together with object=based features from Faster R-CNN [Ren et al. [2017]] models.

In Table 4.1, we show the performance of our Object-Less models (OL–LXMERT) in comparison to the existing state-of-the-art in text classification and multi-modal approaches. Our model obtains promising results against the competitive baselines in both single and multiple modalities. With the same architecture of LXMERT, our Object-less model is more efficient and encouraging to overcome the challenges in images from blind people. Instead of depending on external Object Detection models, our OL–LXMERT is

ideal enough to take advantage of visual features to generate the local and global object representation. In addition, these results also reveal that Visual Question Classification is challenging and distinctive. Without the contribution of image (I), VQC is indeed arduous in text classification. This task is fully worthy enough to be considered in the independent problem in the multi-modal area.

### 4.4.3 Ablation Studies

In this part, we also conduct some ablation studies to emphasize the contribution of our proposed components. Firstly, we also present the reason that we consider Visual Question Classification as an independent task. With a cursory glance, VQC is quite similar to the text classification task in NLP. The results in Table 4.2, however, reflect this task's differences and challenges. In text classification, the category is determined by the context of natural language. However, the type of visual question is based on the relationship between images and texts. Especially, in VizWiz-VQC, questions are spoken by the blind in their daily lives, so it contains a lot of redundant information. In Table 4.2, the single modality model can not overcome the challenges in the VQC task. With the combination of images and questions, our approach and previous multi-modal system have enough features to give a correct choice.

Table 4.2: The contribution of images and texts in multi-modal VQC task

| | **Model** | **Precision** | **Recall** | **F1** |
|---|---|---|---|---|
| Q | BERT | 0.58 | 0.63 | 0.59 |
| I | Vision Transformer | 0.56 | 0.57 | 0.55 |
| QI | OL–LXMERT | **0.67** | **0.69** | **0.68** |

Secondly, we also emphasize the strength of visual features from Transformer architecture against traditional Convolution Neural Work models. In this comparison, we utilize the image features from the latest improvement of the CNN-based model as EfficientNet [Tan and Le [2019]]. Our detailed comparison between Transformer-based and CNN-based image feature extraction is presented in Table 4.3. Obviously, Vision Trans-

former obtains the best performance against the traditional approaches in CNN.

Table 4.3: The comparison of Transformer-based and CNN-based image feature extraction

| | Model | Precision | Recall | F1 |
|---|---|---|---|---|
| **Object-based** | **Faster-RCNN** [Ren et al. [2017]] | 0.63 | **0.70** | 0.66 |
| **Objectless-based** | **EfficientNet-b7** [Tan and Le [2019]] | 0.58 | 0.63 | 0.59 |
| | **EfficientNet-b6** [Tan and Le [2019]] | 0.56 | 0.57 | 0.55 |
| | **ViT (B_16)** | **0.67** | 0.69 | **0.68** |

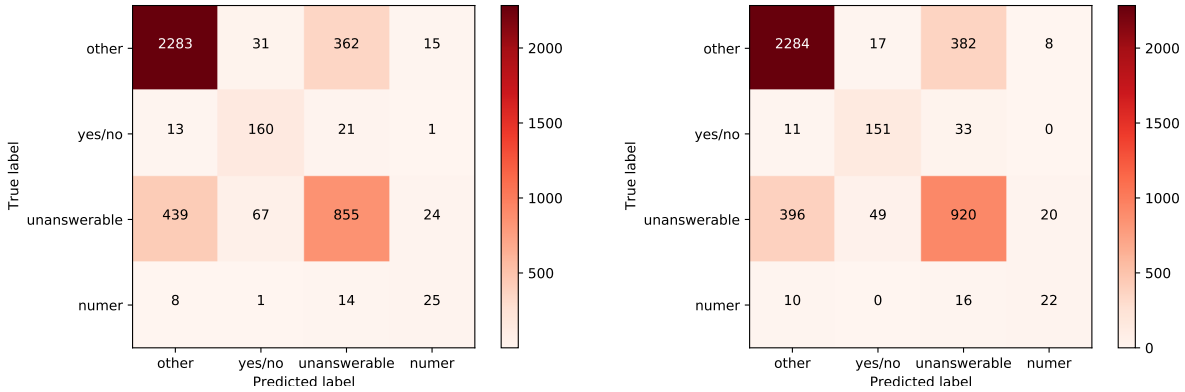
The visible question in this comparison is the weakness of virtual objects in CNN-based approaches against the Object-based model of Faster R-CNN. However, when we consider the architecture of Faster R-CNN carefully, it is easy to realize that both EfficientNet and Faster-RCNN are also deployed by the CNN models. Moreover, the annotated information of Object Detection is more greatly enriched than image classification. Besides, LXMERT [Tan and Bansal [2019a]] architecture is designed and trained to satisfy the object-based models. Therefore, with the same fundamental architecture from CNN, the performance of Faster RCNN is better than EfficientNet in the LXMERT model.

However, as we mentioned above, the drawback of Convolution Neural Network models is based on the mechanism of kernels which only observe limited regions in images. In Object-less images, there are a few objects in high confidence, which means the features of local objects are less meaningful to cover all content of images. On the contrary, our model takes advantage of the Transformer-based approach to extract the global features in the object-less images. With the same process of virtual objects, the global features from Transformer architecture obtain significant results against CNN-based representation in both Image Classification and Object Detection models. It also reflects that our virtual images are auspicious enough to integrate the global and local features in images.

## 4.5 Discussion

As a result, our virtual objects are created to reduce the dependence on object-based features in vision-language models. However, there are some special cases where the content of objects is so important. Through the detailed confusion matrix in Figure 4.5,

the strength and weaknesses of our virtual objects are demonstrated clearly.




(a) LXMERT (b) OL–LXMERT

Figure 4.5: The detailed confusion matrix between LXMERT and OL–LXMERT

Firstly, our Object-less LXMERT model outperforms the object-based LXMERT in total. Specifically, our model has significant success in the unanswerable samples which are the characteristic features of the VizWiz dataset against the previous works such as VQAv2.0 [Goyal et al. [2017]], CLEVR [Johnson et al. [2017b]], etc. The main reason for the appearance of unanswerable samples in VizWiz comes from the collection process from blind people. Obviously, these samples are too poor-qualified to detect any objects in images. Therefore, in these cases, our virtual objects are indeed efficient to predict the category of visual questions.

With the *other* question, both LXMERT and OL-LXMERT are equivalent in precision, recall and F1-score. Nevertheless, the recall of OL-LXMERT is worse than LXMERT model in *number* and *yes/no* question. It means that the object-based model focuses on the relationship between objects and keywords in question to predict the type of question. It also comes from the characteristic of the *number* and *yes/no* visual samples whose content is about the existence and quantity of objects. However, the precision of the LXMERT model in these kinds of questions is worse than OL-LXMERT. Although these samples tend to relate to the appearance of objects, most images in the VizWiz dataset are object-less. Therefore, our object-less model is more powerful to cover the global and local features through virtual objects.

Obviously, our virtual objects prove their strength and robustness in four kinds of

questions. Especially, in unanswerable questions, our virtual objects clearly prove their importance and potential. Even, in the mainstream of object-based models in number and yes/no visual question, our object-less also obtains significant precision against LXMERT. However, we can not deny that the low recall of OL-LXMERT in the object-based question is also the weakness of our virtual objects. It encourages us to deploy the delicacy combination between real and virtual objects in future works.

## 4.6 Summary

In this work, we emphasize the importance and necessity of the Visual Question Classification task as an independent problem in the multi-modal area, especially in object-less images for blind people. We also propose the Object-Less LXMERT model to take advantage of the pre-trained vision-language model with the fewest modification via transfer learning. Our OL-LXMERT model is efficient to generate the virtual objects for replacing the role of the external Object Detection models in previous vision-language approaches. Through our detailed comparison and ablation studies, our Object-less LXMERT model achieves significant results against the competitive baselines in both single and multiple modalities in our extracted VizWiz-VQC 2020.

# Chapter 5

# Yes/No Visual Question Answering

## 5.1 Introduction

Humans produce a huge amount of data to support our modern life every day. Nowadays, data is in multi-media that often contains both images and texts together. It requires multi-modal approaches to utilize their knowledge and relationship. Therefore, cutting-edge studies in image and text gain more and more interest from researchers. It comes from a close connection between textual and visual objects in most applications. The relationship between images and texts is useful to enhance the knowledge of many tasks and applications. For example, texts reveal the content of images clearly and vice versa. Those tasks, however, seem more and more challenging and difficult due to their variety and diversity. Researchers need to take advantage of both Computer Vision (CV) and Natural Language Processing (NLP) in those tasks. Despite their difficulty, this trend seems practical and attractive in both research and industry. There are a lot of works to deal with image and text challenges like Image Captioning [Yun et al. [2019], Sharma et al. [2018]], Visual Question Answering [Su et al. [2020a], Tan and Bansal [2019b], Hudson and Manning [2019]], Visual Commonsense Reasoning [Wang et al. [2020c], Zellers et al. [2019]], etc. In those studies, the key goal is deriving the novel and profound features from both text and image.

Particularly, Visual Question Answering (VQA) gets a significant concern in both prac-

tical and research. VQA determines a textual answer from an image and an input question. Those answers belong to an object, color, and a group of details in that image. Manually, it is easy to extract the keywords of questions related to the visual content of images. This information of keywords and visual features is useful to predict a specific answer. Unfortunately, the process that includes feature extraction, relationship combination, and answer prediction should achieve automatically. Therefore, the objection of the VQA is to exploit and combine both textual and visual features to find out the answer.

Another problem is whether this task should be a classification or generation task. Like a human's behavior, the generation is similar to our brain's process. Through digesting all the information of both images and questions, we generate the sentence to answer the input. Despite the conventional sense in the generation task, most researchers struggle in a classification task to design their system due to its advantages. In the classification task, VQA is more efficient and easy to deploy in an application. Besides, in most VQA datasets, answers are short and replicated, which is more suitable for the classification task.

Recently, previous approaches consider the answer as the set of potential words and phrases. The length of vocabulary is the key component to gain the success of the system's performance. As a human, we expect a system that can effectively answer all types of questions. Regrettably, no one can know everything. Therefore, it is too fantastic to build a system having unlimited answer vocabulary. We can not expect a system that can answer every kind of question. Concerning these reasons, we propose to narrow down this task into a specific question type. Specifically, in a range of our works, we only focus on Yes/No questions. In most datasets in the VQA task, questions include Yes/No, Number, and Other. Besides, for example, in an image retrieval system, the most popular question is to find the existence of images and objects. It proves that Yes/No questions obtain the concern in both research and application.

Generally, most previous approaches are a combination of question and image understanding components. Specifically, traditional models use a Convolution Neural Network to extract the visual features. Furthermore, understanding questions is done by a Re-

current Neural Network and its variants including Long Short-term Memory [Yang et al. [2016]], Gated Recurrent Unit [Ren et al. [2015a]], etc. It requires huge parameters to learn the feature extractor of the image and question. Recently, in Computer Vision, the idea of deploying powerful and portable modules has arisen rapidly. Particularly, pre-trained models are effective to integrate into the downstream tasks. Therefore, we propose to take advantage of pre-trained models built on the huge dataset instead of developing from scratch. Specifically, we propose to apply delicately pre-trained models to extract textual and visual features. Our question embedding sub-module takes advantage of the BERT model [Devlin et al. [2019]] that is a robust language model built on Transformer architecture [Vaswani et al. [2017]]. Its advantages come from the significant success in many NLP tasks. In image embedding, we make use of two kinds of pre-trained image classification models that include ResNet [He et al. [2016]] and VGG [Simonyan and Zisserman [2014]]. Through their different architecture, ResNet and VGG extract the global and residual features. The variety in our image understanding obtains more efficiency and robustness than a single module.

The other interest of our proposed approach is the combination of image and text. A question only refers to specific objects and regions of images. Therefore, it is an undeniable fact that attention is necessary to capture the relationship between partial image features and textual keywords. In this work, we propose to apply Stacked Attention that relies on an interaction between regional image features and textual question representation. Unlike the previous attention based on the image vectors, Stacked Attention works with region features extracted by Convolution layers. Therefore, our attention is useful to improve the relationship between partial images' regions and textual questions. This attention also contributes to image understanding by integrating textual information into the image's pixels. In this work, we propose BERT-RG, the compact and efficient approach for predicting a visual-textual answer in the Yes/No question type. We also present a novel perspective to deploy a specific system instead of gaining an unlimited and unrealistic model. Furthermore, our model can enhance the regional images related to questions through the Stacked Attention mechanism. Our pre-trained model integration

is compact and efficient due to taking advantage of the external datasets. We conduct all experiments in VizWiz-VQA, a humane and reliable benchmark dataset for blind people. Experimental results show the effectiveness and robustness of our BERT-RG model when compared with competitive baselines.

Our main contributions are as follows:

- Through our careful observation, we propose a novel viewpoint in the VQA problem. Because each system has its domain, a specific approach to deal with one kind of question is more reasonable and effective.
- By our effort to create a compact VQA model, we propose to apply pre-trained models into our architecture. With our delicate combination, these modules are useful enough to decline huge parameters in our model. It is ideal for a practical application.
- To the best of our knowledge, we propose an effectively completed VQA model that outperforms the previous approaches in VizWiz-VQA, the latest and reliable benchmark dataset.
- Specifically, we propose to combine different kinds of visual features effectively. The most strength of our visual features' extractor is so compact and sufficient that it reflects a local and global consideration on images through ResNet [He et al. [2016]] and VGG [Simonyan and Zisserman [2014]] models.

## 5.2 Related Works

Traditional VQA techniques are a combination of recurrent and convolutional neural networks. Firstly, most approaches use a recurrent network [Ren et al. [2015b]] such as Long Short-term Memory (LSTM), Gated Recurrent Unit (GRU) to extract the textual features. Secondly, images are embedded by convolution neural networks [Lu et al. [2015]]. Finally, it uses vector operations such as point-wise multiplication, concatenation, and so

on to combine textual and visual features for answer prediction [Kolling et al. [2020]]. These kinds of approaches are so popular and effective that it is easy to deploy in many systems. However, baseline models built on recurrent and convolutional networks are too hard to obtain significant results. It requires a great effort to optimize models from scratch. The deeper model is, the more time and effort it spends.

In both Computer Vision (CV) and Natural Language Processing (NLP), researchers always intensify their efforts to develop basic and recyclable systems. It has arisen an era of integrating pre-trained models into different tasks. These recent works also realize the importance of pre-trained models in developing a VQA system. A famous and popular model of this kind is FT-VQA [Kazemi and Elqursh [2017]], a strong baseline for the VQA task. It uses a pre-trained convolutional neural network, ResNet-152 [He et al. [2016]], to embed an input's image. ResNet [He et al. [2016]] is based on residual network architecture to obtain high-level features of images. Traditionally, FT-VQA also uses LSTM to embed questions through different sizes in a word embedding. This work indicates the importance of stacked attention in combining image glimpses and textual features. As a result of ResNet integration, FT-VQA [Kazemi and Elqursh [2017]] outperforms the existing methods at that time. The reason for this success comes from the strength of ResNet. A closely related point that deserves attention is the dataset. ResNet and FT-Work use the same dataset, MS COCO dataset [Lin et al. [2014]], for learning. FT-VQA [Kazemi and Elqursh [2017]] takes advantage of pre-trained models to extract the visual features.

Visual attention mechanisms learn to focus on image regions that are relevant to the task. In VQA, previous works also put their concern on the attention of combining the relationship between textual and visual features. Up-Down [Anderson et al. [2018a]] model is a featured system that proposes a novel attention mechanism in image understanding. Generally, there are two kinds of attention that include detection proposals and global attention. Up-Down [Anderson et al. [2018a]] models combine the two approaches into one by generating the global attention map over the local proposals. Specifically, Up-Down [Anderson et al. [2018a]] utilize the pre-trained object detection model - Faster

R-CNN Ren et al. [2017] - to predict regional object proposals. In the Bottom-Up phase, object proposals are mapped into feature space by region of interest (RoI) pooling. Based on intersection-over-union (IoU) thresholds in Faster R-CNNRen et al. [2017], each proposal is filtered, pooled by mean operation into a 2048-d feature map. Then, these pooled feature maps are averaged into a single feature map and fed into the attention LSTM. The output of the attention LSTM is a weight vector that reflects the importance score of proposals. Finally, the attended feature map is calculated by summing all of the pooled feature maps according to their weight vector of LSTM. Questions in VQA are only related to small regions in an image. Therefore, approaches based on object detection models are promising in this sense. However, to achieve the strength of pre-trained object detection models, the quality of images and resources play an important role in the system's performance. If images are blurry and contain a few objects, a detector can not find the suitable regions exceeding the confidence score. On the other hand, if the image resource has no tagged objects, object detection cannot fine-tune effectively.

## 5.3 Methodology

By taking advantage of pre-trained models, our architecture is effective and powerful enough to capture valuable features via fine-tuning techniques. Firstly, we propose to combine the strength of image embedding and the language model with stacked attention. In particular, we apply ResNet [He et al. [2016]] and VGG [Simonyan and Zisserman [2014]] to obtain visual features from the image. Furthermore, we also propose to separate the pre-trained image models from our whole architecture. This idea is novel enough to make our system compact and portable in VQA tasks for Yes-No questions. Secondly, our question embedding utilizes a pre-trained language model, BERT [Devlin et al. [2019]], to extract textual features. The strength of pre-trained BERT is meaningful to reveal the relationship between words and questions. Besides, we propose to exploit the stacked attention mechanism [Yang et al. [2016]] to enhance visual and textual representations. By revealing the relationship between partial and linguistic features, this attention is

useful to map the image's regions into the question's keywords. Based on our proposals, our architecture is more effective and compact than previous systems. It comes from our delicate integration and consideration.

The rest of this chapter is organized as follows:

- Image Embedding (Section 5.3.1) provides the content of two kinds of image models for visual feature extractor.
- Stacked Attention Mechanism (Section 5.3.2) presents the approach to combine both textual and visual features into the VQA model.
- Visual Question Answering (Section 5.3.3) shows the architecture of the independent VQA model. Due to our reasonable and novel proposal, VQA also obtains significant results against the previous ones.

### 5.3.1 Image Embedding

Image feature extraction involves deriving a higher level of information from raw pixel values. In VQA, this sub-module increases the volume of architecture and computational cost for running. In recent approaches, Convolution Neural Network is dominating in image embedding. It requires a huge of time and effort to optimize this phase. Therefore, we propose to apply pre-trained networks to extract the image features. Besides, in our model, we also regard image extraction as an independent phase. We reduce the image embedding's weights from our whole architecture. It is the reason that our model is more flexible and compact to convert images into feature vectors on the practical side.

Although there are a lot of well-known systems in Computer Vision, we propose to integrate two kinds of models that include ResNet [He et al. [2016]] and VGG [Simonyan and Zisserman [2014]]. These models are in image classification tasks and work on the ImageNet dataset [Deng et al. [2009]]. The reason for our choice comes from our observation in the VizWiz dataset [Gurari et al. [2018, 2019]]. Firstly, as we mentioned above, human often finds the detail of the image from the question's context to find an answer.

The answer usually belongs to the object and its content. Therefore, it is simple to conclude that object detection models are more suitable than image classification systems. However, the VizWiz dataset is collected by the blind. Most images contain many noises such as blur, rotation. Undoubtedly, it reduces the quality of the image significantly. In the worst cases, human eyes can not even recognize images' content. Besides, the VizWiz dataset contains no tagged samples for fine-tuning object detection. Therefore, image detection models are not suitable for this dataset. We can not utilize the VizWiz-VQA dataset to intensify the performance of image models. To overcome these difficulties in the VizWiz dataset, we propose to combine two types of image classification models to extract visual features. An interaction between ResNet and VGG enhances the image embedding phase instead of applying each model.

### VGG

VGG is published in 2014 by Simonyan and Zisserman [2014]. It is one of the most popular Convolution Neural Network architectures in the Imagenet dataset. The success of the VGG network comes from the development of deeper networks with much smaller filters. Nevertheless, it solves the explosion of parameters in the Convolution layers and enhances training time by replacing large kernel-sized filters with multiple small kernel-sized filters one after another. The VGG model is a stack of convolutional layers with small kernels. It produces a large-sized model that takes so much time to train. There are multiple variants of VGG responding to the total number of layers in the network. In our model, we use the pre-trained VGG16 model that has 16 weight layers in the network. With a large number of parameters, this network can capture highly detailed features via its depth. Therefore, features from VGG contain the global information extracted through the convolution operation.

### ResNet

The depth of CNN is related to the performance of systems. However, the explosion of parameters and training time is an issue. Besides, a vanishing gradient occurs through the

increased depth. Unlike traditional sequential network architectures, ResNet [He et al. [2016]] relies on micro-architecture modules. ResNet architecture makes use of shortcut connections to overcome the vanishing gradient problem. The success of ResNet comes from the strength of residual blocks repeated throughout the model. It helps ResNets to go deeper while the number of total layers remains. Specifically, in the residual block, identity shortcut connections transmit the signal from one layer to another while skipping in several ones. By residual mechanism, it is useful to learn the interaction of features and decide which is the most meaningful information to remain. The residual blocks can transmit the local information throughout network architecture.

### Image Embedding

Through our observation, we propose to integrate the strength of ResNet and VGG in our architecture. The reason for our choice is the different strengths in the two kinds of image classification models. Specifically, our image representation is a combination of global and residual features from ResNet and VGG. It is important to note that ResNet and VGG are feature extraction modules instead of internal components in our architecture. It means that our model has no increased parameters despite the combination of two huge models. With our novel phase, it is enough compact to put it into practice. By considering image embedding as a separate module, it is easy to replace it with the other ones.

However, these models are trained on Imagenet [Deng et al. [2009]] and are not related to the VizWiz dataset. We apply the fine-tuning technique to evolve image embedding models toward specializing in our tasks. In particular, we change classifier layers of ResNets [He et al. [2016]] and VGG [Simonyan and Zisserman [2014]] by Fully-connected Neural Networks (FC) whose output nodes are equal to the size of answer vocabulary in our task. Visual features come from inputs of Adaptive Average Pooling layers. The detail of the architecture is presented in Figure 5.1. Undoubtedly, these features from VGG and Resnet contain the global and residual representation of images.

Although massive effort has been directed toward optimizing the image embedding

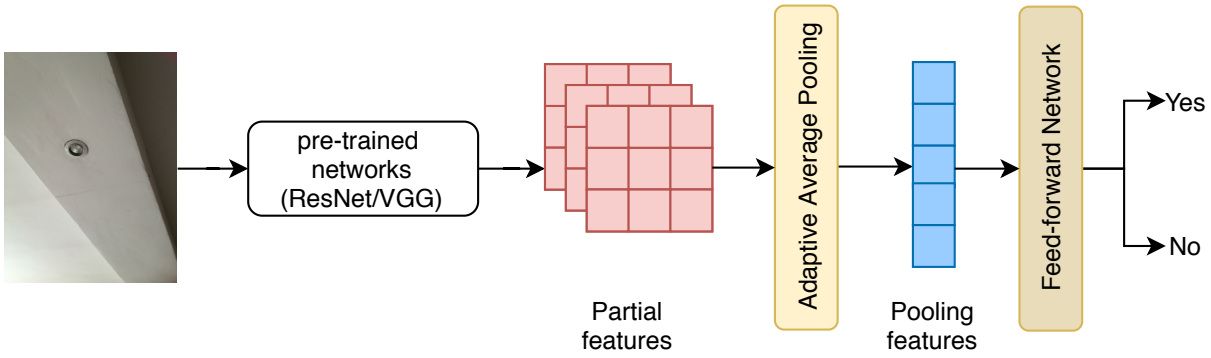



Figure 5.1: Image Embedding: To enhance the relationship between pre-trained image models and our configuration, we fine-tune them with a modified VizWiz-VQA dataset. Specifically, we add a fully-connect layer to classify an image into VQA answers. In the image extraction phase, we utilize the partial features of the Adaptive Average Pooling layer's input.

in the VQA task, we emphasize that this phase is independent of our systems. It is regarded as the automatic feature extractor instead of the handcrafted solution. Each image contains two visual representations from the fine-tuning ResNet and VGG models.

### 5.3.2 Stacked Attention Mechanism

As we mentioned in previous parts, all samples in the VizWiz dataset are collected by blind people. Therefore, the quality of images is not good enough for a traditional convolution neural network. We need more effort to extract the image features. As a result, we propose to combine two kinds of image architectures, ResNet and VGG. Our proposal inherits the strength of two image models to extract both residual and global features. However, the essential question is how to combine textual and visual representation. In the VQA task, answers and small regions of images always exist in the correlation. It means there are more and less attentive parts of images. Therefore, considering an image in a 1-D vector is less smooth and efficient than regional features. These flattened features of images are often too general to map in the question's keywords. Indeed, on manual processing, we always focus on a specific region to answer the VQA question. There was no doubt about the importance of regional information in images. Through our observation, we integrate the Stacked Attention Mechanism into our model to enhance the regional image features

via textual questions.

Mathematically, regions have a score that represents their correlation with textual features. Our model learns and optimizes an attention scoring function between regional and textual information of questions and images. Specifically, we inherit the advantages of the stacked attention mechanism. Our attention filters out noises from unrelated regions and emphasizes the attentive ones. The higher attention score reflects the relevance between image regions and questions.

The detail of stacked attention for one region vector is visualized in Figure 5.2. An image is extracted to visual features by ResNet and VGG. Instead of considering the last layer's output, we obtain the partial features from the previous layer of the adaptive average pooling. Next, we calculate the attentive score of regional and textual vectors. We also enhance the visual vector with the attention signals in the previous step. Finally, we combine the intensive image and question to generate the new query's image.

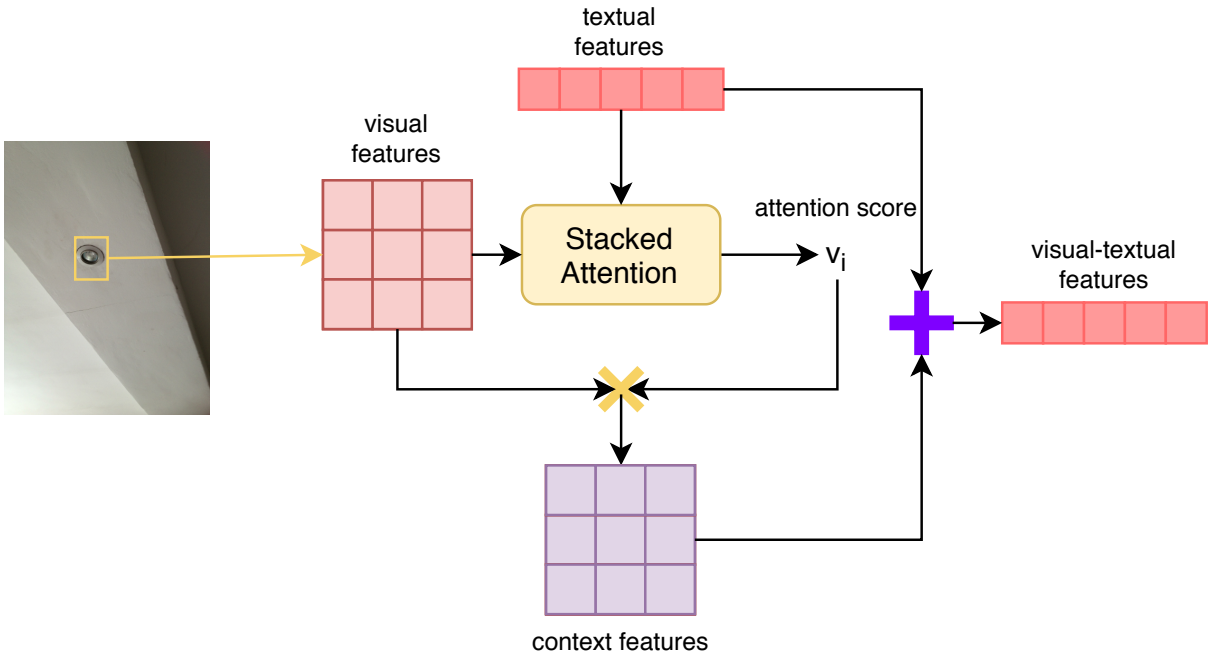



Figure 5.2: Stacked Attention for combining visual and textual features

In particular, we only consider stacked attention in the general image extractor instead of both the ResNet [He et al. [2016]] and VGG [Simonyan and Zisserman [2014]] model. For implementation, the process is similar to the general one. We also execute some

modifications to fit the specific model. For an image $I \in \mathbf{R}^{d_1 \times d_2}$, its visual features are extracted by image embedding $\Gamma_I$ in Equation 5.1.

$$f_I = \Gamma_I (I) \tag{5.1}$$

where $f_I \in \mathbf{R}^{d \times m}$, $d$ is image feature's dimension and $m$ is the number of regions. Similarly, a corresponding question $Q$ is also mapped into textual space by Question Embedding $\Gamma_Q$ in Equation 5.2.

$$f_Q = \Gamma_I (Q) \tag{5.2}$$

Then, the attention distribution of regions is scored through the stack of single Full-connected Neural layer (FCL) between visual features $f_I \in \mathbf{R}^{d \times m}$ and textual vector $f_Q \in \mathbf{R}^d$ in Equation 5.3. These scores are normalized by softmax function over the regions of image in Equation 5.4.

$$h_a = \tanh (W_{IA} f_I \oplus (W_{QA} f_Q + b_A)) \tag{5.3}$$

$$p_I = softmax (W_p h_A + b_p) \tag{5.4}$$

where $W_{IA}, W_{QA}, W_p$ are learning parameters of FCL. The attention distribution $p_I \in \mathbf{R}^m$ reflects the importance score of m regions in image features. Besides, we consider $\oplus$ as the addition of a matrix and a vector. In particular, the normalized image features $W_{I,A} f_I \in \mathbf{R}^{d \times m}$ are regarded to be $m$ vectors that represent $m$ regions in image. Therefore, the addition function $\oplus$ is calculated by adding column of matrix $W_{I,A} f_I$ by question vector.

Next, a contextual image representation is a sum of multiplication between attention scores and partial image features in Equation 5.5.

$$\hat{f}_I = \sum_i^m p_i f_{I_i} \tag{5.5}$$

Where $f_{I_i}$ is the column of image feature $f_I$. After, the refined query vector $f_{iq}$ is the sum of context features and textual information from original question in Equation 5.6.

$$f_{iq} = \hat{f}_I + f_Q \tag{5.6}$$

Generally, we present all the above equations in one layer of attention. However, questions are too complicated to combine with images with only one attention. In Visual Question Answering, images contain a bunch of objects which are related to many question keywords. Therefore, stacked attention is an iteration of the above attention process by applying multiple layers through the enhanced query vector $f_{iq}$. In particular, our model works on k-layers of visual attention as follows:

$$h_a^k = \tanh\left(W_{IA}^k f_I \oplus \left(W_{QA}^k f_{iq}^{k-1} + b_A^k\right)\right) \tag{5.7}$$

$$p_I^k = softmax\left(W_p^k h_A^k + b_p^k\right) \tag{5.8}$$

where $f_{iq}^0$ is initialized to be $f_Q$. In each layer, the refined query vector is intensified by the previous states in Equation 5.9.

$$f_{iq}^k = \hat{f}_I^k + f_{iq}^{k-1} \tag{5.9}$$

The context representation of visual features and attention score is in Equation 5.10.

$$\hat{f}_I^k = \sum_i^m p_i^k f_{I_i}^k \tag{5.10}$$

On behalf of visual and textual combination, a highly refined query vector reflects the relationship between image and question. With multiple layers of attention, questions are digested acutely via many regions in the images.

### 5.3.3 Visual Question Answering

In our VQA model, we propose to utilize the strength of pre-trained models to reduce the system's volume and enhance the prediction performance. We visualize the detail of our architecture in Figure 5.3. Our system consists of three components as follows: (i) image embedding; (ii) question embedding (iii) stacked attention and classifier.

In particular, our question embedding is similar to Section 3.2.1. Furthermore, an input image is extracted by two kinds of image embedding that include ResNet [He et al. [2016]] and VGG [Simonyan and Zisserman [2014]]. We also note that these embeddings are fine-tuned in Figure 5.1 and regarded as the external components in our completed model. It means that the parameters in ResNet and VGG are independent of our system. In the visualization, we only use the partial features in the previous layer of adaptive pooling. The difference in architecture between ResNet and VGG reflects two kinds of consideration in images. ResNet focuses on the residual and local information while VGG generates the global features by multiple convolution layers. However, the large size of ResNet and VGG is a limit to deploying the VQA system. Therefore, we propose to eliminate them from our architecture and utilize it as the feature extractor module. Besides, we also fine-tune them in the image classification task from our VQA model. It means our visual features further relate to the answer in the VQA task.

On the other hand, questions are embedded by the question embedding in Section 3.2.1. Unlike image feature extractors, this embedding is an internal component in our model. After obtaining the textual features, we take advantage of stacked attention in Section 5.3.2 to find the combination of partial features and input questions. Mathematically, the image features $f_r, f_v$ from ResNet and VGG embeddings are combined with textual features $f_q$ from Question Embedding (Section 3.2.1) by stacked attention $\Psi(.,.)$ in Equation 5.13. In our model, we also use the k-multiple attention with a value of $\Psi(.,.)$ is from the last k-th layer of stacked attention.

$$f_r = ResNet(I); f_v = VGG(I) \tag{5.11}$$

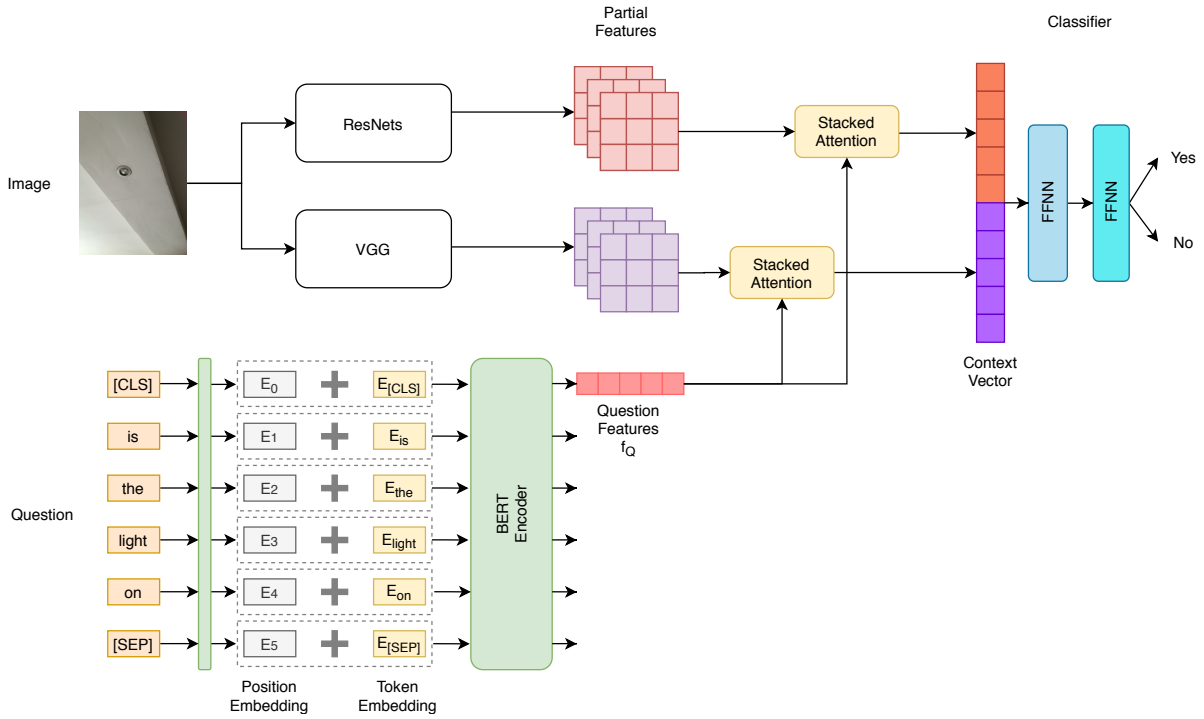



Figure 5.3: Our VQA architecture: The image is extracted by pre-trained ResNet and VGG while the question is embedded via BERT networks. We apply the stacked attention mechanism for each pair to combine textual and visual features. The final representation is a concatenation of outputs between ResNet and VGG branches. For classifier, we use two Feed-forward Networks.

$$f_q = QstEmbedding(Q) \tag{5.12}$$

$$f_{rq} = \Psi(f_r, f_q); f_{vq} = \Psi(f_v, f_q) \tag{5.13}$$

Where $f_{rq}, f_{vq} \in \mathbf{R}^d$. After, these features are concatenated to obtain the context vector which is put into two layers of Feed-forward Neural Network (FFNN) as a classifier in Equation 5.14.

$$p = \sigma\left(W_{c_2}\left[ReLU(W_{c_1}[f_{rq}||f_{vq}] + b_{c_1})\right] + b_{c_2}\right) \tag{5.14}$$

Where $W_{c_1} \in \mathbf{R}^{2d \times h}, W_{c_2} \in \mathbf{R}^{h \times |vocab|}, b_{c_1}, b_{c_2}$ are the FFNN's parameters and $|vocab|$ is the number of answers in our model.

## 5.4 Experiments

### 5.4.1 Dataset

Throughout our works, we conduct all experiments in VizWiz-VQA 2020 dataset. In this task, we only put our concentration on the Yes/No question type. Therefore, we extract this kind of question from the whole dataset. The detail of the Yes/No dataset is presented in Table 5.1.

Table 5.1: The detail of Yes/No question in VizWiz-VQA 2020

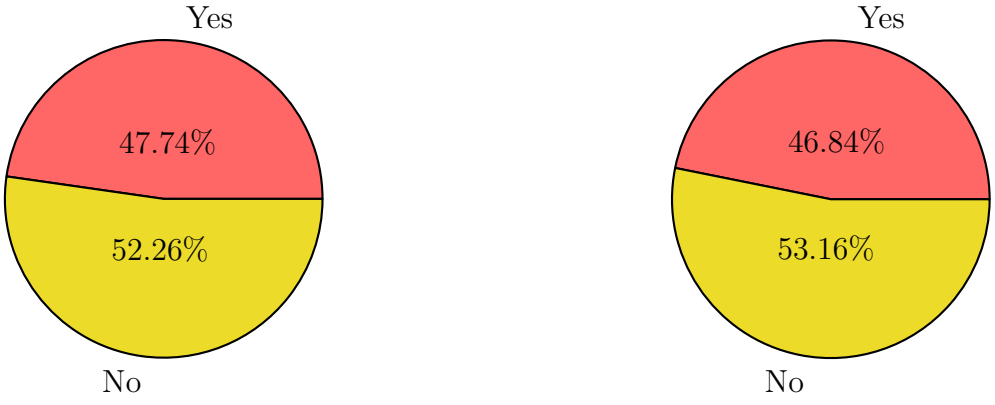
| | **Train** | **Validation** |
|---|---|---|
| **No. Question** | 957 | 195 |
| **No. Answer** | 9570 | 1950 |
| **Size. Answer Vocab (not-preprocessing)** | 830 | 182 (OOV: 36) |
| **Size. Answer Vocab (preprocessing)** | 2 | 2 |

However, in our model, we consider the VQA task in binary classification. We need to select a specific label for one sample. Therefore, we choose the dominating answer in 10 manual ones as the gold label for training. Therefore, after our pre-processing, the answer vocabulary only includes $\{yes; no\}$.

After our configuration, the gold label's distribution in the train and validation dataset is presented in Figure 5.4. This balanced percentage of these labels reflects that this dataset is quite suitable for the Yes/No classification. Therefore, there is no doubt about the importance and necessity of the Yes/No VQA task.

### 5.4.2 Evaluation

In most VQA datasets, each sample contains one image, question, and ten answers from annotators. In the VQA task's configuration, the most popular evaluation metric is accuracy. This accuracy is calculated by the overlap between predicted and gold answers in Equation 5.15.




(a) Train (b) Validation

Figure 5.4: Label Distribution of VizWiz-VQA dataset in Yes/No question

$$acc(predict) = \min\left(\frac{\sum_{i=1}^{10} ||predict = answer_i||}{3}, 1\right) \tag{5.15}$$

where $||.||$ is a boolean function to determine whether a system's answer is similar to manual one or not. Generally, an answer is considered as correct one if it is agreed by at least 3 people.

In a testing phase, we make use of the online platform of VizWiz VQA Challenge 2020. All results are extracted from EvalAI[1] in "test-standard-2020" phase. From the content of the VizWiz-VQA challenge, the results come from VQA accuracy. Our results appear on the leaderboard of VizWiz challenge 2020.

### 5.4.3 Implementation Details

The details of our configuration are presented in Table 5.2. Specifically, we initialize our parameters from the pre-trained models. Firstly, in image embedding, ResNet-152 and VGG-16 are integrated into our networks in Figure 5.1. After that, we enhance our image extractor by fine-tuning with the answer from the VizWiz-VQA dataset. Secondly, in question embedding, we initialize our model with the pre-trained weights from BERT [Devlin et al. [2019]] in the settings. We are also concerned about the effect of non-contextual word embeddings in question understanding. Specifically, we also use the

[1] https://eval.ai/web/challenges/challenge-page/743/overview

pre-trained weights of Glove [2] and Word2Vec[3]. We also present the detail of settings in Table 5.2.

Table 5.2: The detail of our models in experiment and ablation studies

| Component | Value |
|---|---|
| ResNet | 512 |
| VGG | 16 |
| BERT | 768 |
| Hidden Size | 512 |
| FFNN-Classifier | 1024/512 - 170 - 2 |
| Word2Vec | GoogleNews-vectors-negative - 300 |
| Glove | glove.840B - 300 |
| LSTM | hidden =1024;layer = 1 |
| CNN | stride=1,filter=512,kernel =[2,3,4,5] |
| m (regions) | 7 |
| Activation | Tanh, ReLU, Sigmoid |

### 5.4.4 Experimental Results

Table 5.3 shows the experimental results of our model against the state-of-the-art models in the VizWiz-VQA dataset. In this comparison, BERT-RG is our completed model while FT-VQA [Kazemi and Elqursh [2017]] and Up-Down [Anderson et al. [2018a]] are the existing baselines in many VQA datasets. All results in Table 5.3 comes from the online evaluation.

Table 5.3: The experiments' comparison in Yes-No Questions set of VizWiz dataset

| Methods | VQA Accuracy |
|---|---|
| **FT-VQA [Kazemi and Elqursh [2017]]** | 68.10 |
| **Up-Down [Anderson et al. [2018a]]** | 59.60 |
| **Katya[1] Leaderboard-2020** | 77.84 |
| **HSSLab[2]Leaderboard-2020** | 78.89 |
| **VT-Transformer** | 73.64 |
| **BiCAtt VT-Transformer** | 75.64 |
| **BERT-RG(Our model)** | **79.85** |

FT-VQA [Kazemi and Elqursh [2017]] is trained in VQA-v1.0 [Goyal et al. [2017]] and

[2] https://nlp.stanford.edu/projects/glove/
[3] https://code.google.com/archive/p/word2vec/

fine-tuned in the Vizwiz-VQA dataset. Furthermore, Up-Down [Anderson et al. [2018a]] is trained in the VizWiz dataset from scratch. In our work, we put no concentration on re-implementation. Therefore, FT-VQA and Up-Down results are extracted from VizWiz contest [Gurari et al. [2018]].

Another comparison in this part comes from the leaderboard of VizWiz-VQA challenge 2020. Two of the most significant models are presented in Table 5.3 as Katya and HSS-Lab[4]. These results are reliable and up to date in the VizWiz-VQA dataset. We only consider these results in the Yes/No question type.

In these comparisons, our model, BERT-RG, outperforms all baseline models. Even with our different modules of BERT-ResNet, BERT-VGG, and BERT-RG w/o Attention, our architectures also obtain equivalent results against SOTA models. Unlike the previous approaches, our model focuses on a specific kind of question. It is reasonable in both research and application. With over 10% higher than FT-VQA (published model) and 1% higher than HSSLab (real-world model from VizWiz-VQA 2020 challenge), our model achieves effectiveness through our novel viewpoint in the VQA task.

### Ablation Study

Specifically, we conduct ablation studies to evaluate the contribution of each component in our model. All results in this part are extracted in the EvalAI system from the test-dev dataset. It is different from the test-standard set in Table 5.3. The reason for our configuration comes from the settings of the VizWiz-VQA platform. In the test-standard set, there is only one chance to evaluate per month and less than 5 times. Specifically, in this part, we focus on 5 main points as follows.

- The importance of question and image in VQA
- The strength of our proposed image embedding against the traditional CNN models
- The strength of our proposed consideration in fine-tuning of image embedding

[4] https://eval.ai/web/challenges/challenge-page/743/overview

- The strength of BERT against the traditional approaches in question embedding
- The effect of stacked attention and attention layers in the VQA system

Firstly, we focus on the difficulty of the VQA task for blind people. If we only use either questions or images to answer the sample, the best performances we can get are presented in Table 5.4. Even in the case of using the only image, our proposed image embedding also outperforms independent ResNet and VGG. Obviously, the fully completed combination of question and image is better than the others.

Table 5.4: The performance in the intersection of image and question (test-dev)

| Model | | VQA Accuracy |
|---|---|---|
| **Image** | **VGG16** | 70.60 |
| | **ResNet** | 63.30 |
| | **RG (Ours)** | 72.47 |
| **Question** | **BERT** | 78.65 |
| **Image+Question** | **BERT-RG** | **82.96** |

In Section 5.3.1, we conclude that ResNet and VGG treat an image in completely different ways. Indeed, Table 5.5 presents the results of BERT-ResNet and BERT-VGG. We can easily reveal the difference between ResNet and VGG. Obviously, in the VizWiz-VQA dataset, a variety of image quality has arisen a demand for powerful approaches. Generally, VGG is better than the residual features of ResNet in this dataset. However, the combination of ResNet and VGG proves the importance of both residual and global features in the image. By combining ResNet and VGG into our architecture, BERT-RG is more powerful and effective than a single image embedding. In Section 5.5, we also present examples to prove their contribution to the success of our architecture.

Table 5.5: The comparison between BERT-ResNet and BERT-VGG

| Model | VQA Accuaracy |
|---|---|
| **BERT-ResNet** | 76.41 |
| **BERT-VGG** | 78.41 |
| **BERT-RG** | **79.85** |

As we mention in the Image Embedding part, we train the ResNet and VGG in VizWiz-VQA without the participation of questions. These components are used independently

as the visual feature extractor with the pre-trained weights from our dataset. We also point out that in the small dataset as VizWiz-VQA, the fine-tuning of the image extractor is better than the originally pre-trained VGG and ResNet. Obviously, pre-trained VGG and ResNet are trained on the large image classification dataset which is different from ours. Therefore, the fine-tuning process is essential to adapt the huge model in the small downstream task. In Table 5.6, the performance of BERT-RG in all modes with fine-tuning is better than the original model in image embedding.

Table 5.6: The effect of fine-tuning in image embedding (test-dev)

| Model | Mode | VQA Accuracy |
|---|---|---|
| **W. FineTuning** | **Att. Layer = 1** | 80.71 |
| | **Max (Att.Layer=8)** | 82.96 |
| | **Min (Att.Layer=15)** | 78.46 |
| | **Average** | 80.60 |
| **W/o FineTuning** | **Att. Layer = 1** | 77.72 |
| | **Max (Att.Layer=10)** | 80.52 |
| | **Min (Att.Layer=8)** | 77.15 |
| | **Average** | 78.51 |

Furthermore, the integration of BERT for question embedding is again a strength of our approach. Therefore, we would like to compare our BERT-RG model with two non-contextual embeddings that include Word2Vec [Mikolov et al., 2013] and Glove [Pennington et al., 2014].

Table 5.7: A comparison of sentence embedding

| | Glove | Word2vec |
|---|---|---|
| **CNN** | 77.27 | 73.16 |
| **LSTM** | 64.47 | 63.71 |
| **BERT** | **79.85** | |

To analyze the success of our architecture, we reproduce two popular text representation approaches that include Long Short-term Memory Network [Hochreiter and Schmidhuber, 1997] (LSTM) and Convolution Neural Network [Kim, 2014a] (CNN) whose settings are presented in Table 5.2. In each model, we also use both Glove and Word2Vec in the word embedding layer. Table 5.7 presents our comparison to reveal the effectiveness of BERT against two famous and traditional models. Specifically, we also try many different

settings of LSTM and CNN such as the number of filters, hidden dim. The trend from LSTM and CNN, however, does not change too much.

Table 5.8: The comparison between No Attention and Stacked Attention

| Model | VQA Accuaracy |
|---|---|
| **BERT-RG w/o Attention (BERT-NoAtt)** | 78.20 |
| **BERT-RG w Attention** | **79.85** |

Another characteristic we would like to mention is Stacked Attention. In this comparison, we present the strength of stacked attention in the combination of visual and textual features. Against Stacked Attention, BERT-RG w/o Attention uses vector multiplication to combine question and image. The detail of attention's comparison is shown in Table 5.8.

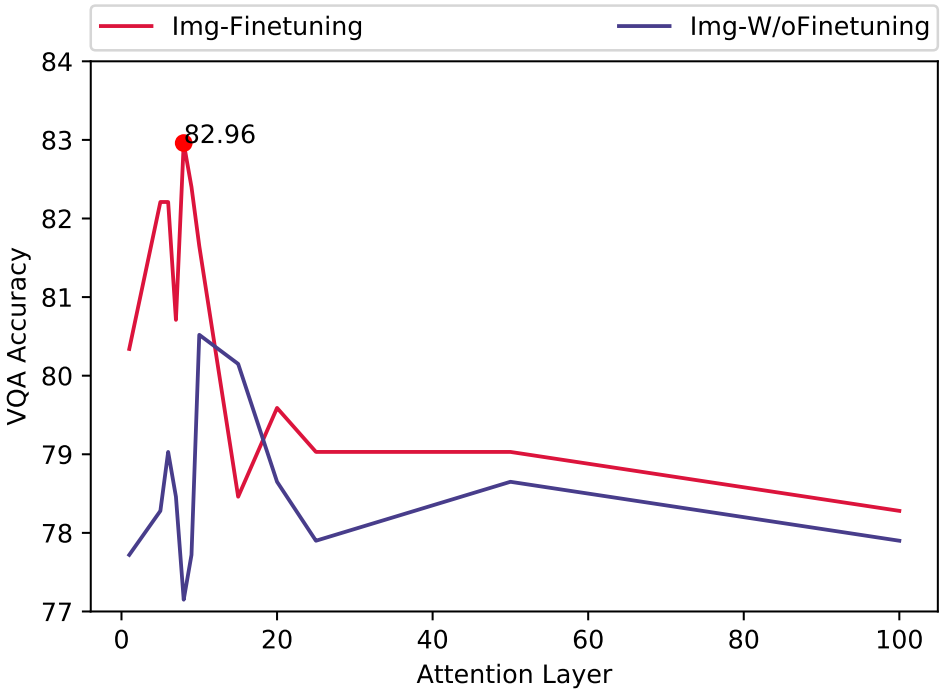



Figure 5.5: Stacked Attention for combining visual and textual features

In the image model's study, we illustrate the difference between ResNet and VGG. It has aroused a question of how to combine them and question. Stacked Attention plays an important role in combining visual and textual features. The attention digests a relationship between image and question to emphasize the meaningful regions and enhance

a query. Unfortunately, this attention is calculated by a combination of different outputs from neural layers, which is too hard to visualize in the original image and question. However, the numerical result efficiently proves the effectiveness of stacked attention in our model.

With the strength of the stacked attention mechanism, we also present the effect of the attention layer in this component. In our comparison, we also compare the strength of fine-tuning against no fine-tuning in the increase of attention layer. All results are visualized in Figure 5.5. Obviously, in the early part of the attention layer, the performance of our models is increased gradually. From the 10th layer in stacked attention, the accuracy decreases a little bit.

## 5.5 Discussion

### 5.5.1 Dataset Challenges

VizWiz-VQA dataset comes from the community's contribution to support disabled people. In this dataset, blind people use their mobile phones to take pictures and record their questions. Therefore, this process also causes many problems in this dataset. Firstly, it is easy to observe the degradation in images and ambiguity in questions. Secondly, when we study the VizWiz Challenge, we can recognize the new problem. In VizWiz Challenge, there are a lot of unanswerable samples to provide for the Answerability prediction task. In short, there are three main challenges in VizWiz-VQA as follows:

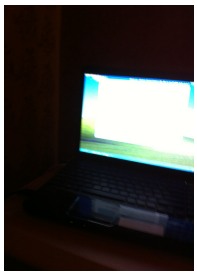
#### Poor Quality of Image

As we mentioned above, blind people take all images by themselves. It is the reason that the quality of images is quite low. Examples of these images are shown in Figure 5.6. In Figure 5.6-($a$), an angle of view does not focus on the object - the piece of mail. Therefore, even humans are not able to detect what it is. Another problem we can easily meet is blurred images such as Figure 5.6-($b, e, f$). Moreover, some images are taken

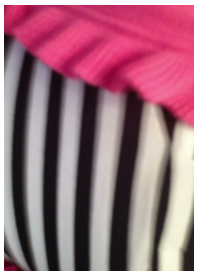
(a) *VizWiz_train_*00000024
Q: Can you see the piece of mail and can you tell me who this mail is for?
A: No

(b) *VizWiz_train_*00000727
Q: Can you read this label?
A: No

(c) *VizWiz_train_*00002023
Q: Hey there, I need to know if this is a bus stop?
A: Yes



(d) *VizWiz_val_*00000919
Q: It is a box.
A: No



(e) *VizWiz_val_*00003550
Q: Does this sweater and this shirt go together?.
A: Yes

(f) *VizWiz_val_*00004140
Q: Is this currently set on the off position?
A: Yes

Figure 5.6: Examples of poor quality images

against the light source. It makes images blurry in the dark background. Figure 5.6-$(c, d)$ has a challenge in learning image's features without human effort. All examples prove the difficulty of the VizWiz-VQA dataset in image understanding. However, the more challenges there are, the more interest researchers obtain.

### Difficult Question

Another challenge in the VQA task is to extract the meaning of the questions. In our work, we put our concentration on the Yes/No question type. The system needs to find out the meaningful keywords related to the image's content.

However, all questions in the VizWiz-VQA dataset come from the daily text of blind people. Therefore, it contains a lot of redundant and ambiguous questions. We also present some examples for these questions in Table 5.9. In example $(1, 2)$, there is much conventional information. It is too hard to extract meaningful keywords. Moreover, the

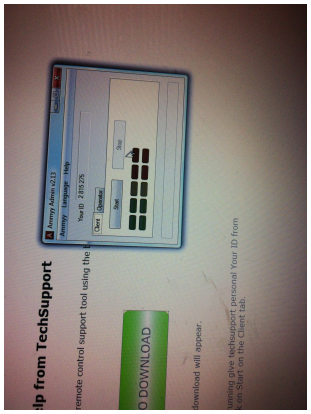
Table 5.9: Examples of Difficult Questions in VizWiz-VQA Dataset

| No. | ID | Question |
|---|---|---|
| **1** | **VizWiz_val_00000319.jpg** | There should be a skin of a document on the screen. Can you please tell me if that's true or not?<br>Thank you very much |
| **2** | **VizWiz_val_00002717.jpg** | Does this outfit go together?<br>Hopefully there's enough light.<br>Thank you. |
| **3** | **VizWiz_val_00001047.jpg** | This is a computer. |
| **4** | **VizWiz_val_00003137.jpg** | Getting closer to the boyle? |

example (3) is an affirmative sentence instead of an interrogative one. Besides, in some cases, the question's words are not correct in the English dictionary.

### Unanswerable Sample

In VizWiz Challenge 2020, there is an independent task of predicting the answerability of a visual question. Therefore, there are a lot of unanswerable samples in the dataset. It causes too much noise for the learning phase. Specifically, gold labels are dominant among ten human answers. In the worst cases, we can not determine which gold label is for training. Figure 5.7 presents the specific example of the unanswerable sample due to the annotators' conflict. In this example, the answer type is answerable, yet it is impossible to determine the answer based on VQA accuracy. It has arisen the importance of a specific question answering system instead of an unlimited approach.




- **Question:** OK, lets try this again and see if this works. Do you see any kind of an ID number here that would tell me what is what so that I can get disconnected?
- **Gold Answers:** {no but youre mouse almost on top stop button (1); id number present (1); unsuitable image(2); unanswerable (2); yes (2); no (2)}

Figure 5.7: An Example of Unanswerable sample in VizWiz-VQA 2020: The inconsistency of annotators leads to no existence of a dominating answer agreed by at least three people.

### 5.5.2 Experimental Analysis

This part provides useful examples to analyze the performance of our models. Through experimental analysis, the strength and weaknesses are easily identifiable in practical examples. The details of these samples are presented in Figure 5.8. Unfortunately, VizWiz-VQA's test set is not allowed to be accessed by researchers because all evaluation comes from the automatic system on EvalAI website[5]. It is useful to protest a secret of the test set and do justice to everyone. Consequently, validation samples are more suitable in this situation.

The first one we would like to mention is 5.8-($a$) whose question is regarded as a difficult one in Table 5.9. This question is too hard to extract the most important keywords related to an image. Therefore, our models can not predict the right answer for this sample. However, the gold answer of this sample is correct. Although this label comes from the annotators, we easily recognize that this image is the photo's skin. It is the reason that our models can not succeed in this case. It also reflects that our models are effective to combine meaningful keywords and images to determine the answer.

The other examples come from the effect of stacked attention in our models. In example 5.8-($b$), both ResNet and VGG can extract the right answer, whereas BERT-NoAtt is not successful in this case. VQA system requires the relationship between image and question. In our model, stacked attention plays an important role in connecting textual and visual features. In this case, despite meaningful image features, the model cannot learn which part in question is related to the image for finding an answer, even this question is obvious to understand. By attentive observation in image 5.8-$b$, there are many noisy objects which are required attention to reveal the specific image's features related to the question's keywords - *the computer screen*. In these cases, stacked attention is significantly useful in the VQA task, which leads to better success in our proposed models.

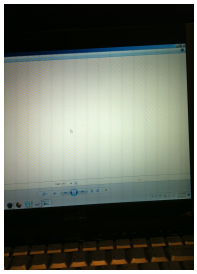
The other aspect we would like to be concern about is the strength of image embedding in the BERT-RG model. In our model, each image embedding plays a different role in

[5]https://evalai.cloudcv.org/web/challenges/challenge-page/523/overview

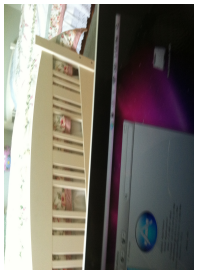
learning images. ResNet can extract the residual and local features while VGG puts its concern on global features. In some cases, the combination of residual and global features is meaningful to find the correct answer. In example 5.8-$(d, f)$, although BERT-ResNet and BERT-VGG predict incorrectly, the BERT-RG model succeeds in utilizing two kinds of features to generate the answers. It proves that the combination of two image embeddings is more robust than a single one. Even our BERT-RG models also eliminate the noise in image embeddings. In example 5.8-$(c, e)$, although either BERT-ResNet or BERT-VGG can answer correctly, BERT-RG is not further degraded by a weak image embedding. Our model, BERT-RG, also predicts the correct answer against the inconsistency of ResNet and VGG. In these cases, our proposed image embedding proves its strength in the success of our model.

## 5.6 Summary

In this work, we introduce a practical viewpoint of VQA to focus on a specific kind of question. Instead of gaining more effort into a comprehensive VQA system, we propose a novel Yes/No VQA system that limits the answer's vocabulary to support and solve the binary questions. Our model takes advantage of two kinds of image embeddings that include ResNet and VGG. While ResNet extracts the residual features, VGG succeeds in revealing the global information of images. We also integrate the most recently popular language model-BERT into our question embedding. Through fine-tuning techniques, our models get the strength of pre-trained models to understand images and questions. Besides, we propose to use Stacked Attention for efficiently combining visual and textual features in our VQA architecture. This attention is also meaningful to find the relationship between images and questions. Our delicate combination is novel to enhance the image and question's understanding. Through the detailed experiments, our model outperforms the existing methods in the Yes/No question. Our success also comes from the scientific research and practical challenge of the VizWiz-VQA dataset. Our extensive analysis of the VizWiz-VQA dataset opens the social interest in research to deploy the



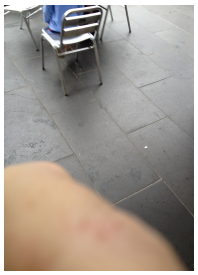
(a) *VizWiz_val_*00000319
Q: There should be a skin of a document on the screen. Can you please tell me if that's true or not? Thank you very much
**Gold**: Yes
**BERT-RG**: No
**BERT-ResNet**: No
**BERT-VGG**: No
**BERT-NoAtt**: No



(b) *VizWiz_val_*00000333
Q: Does my computer screen need cleaning?
**Gold**: Yes
**BERT-RG**: Yes
**BERT-ResNet**: Yes
**BERT-VGG**: Yes
**BERT-NoAtt**: No

(c) *VizWiz_train_*00000444
Q: Is this currently set on the off position?
**Gold**: Yes
**BERT-RG**: Yes
**BERT-ResNet**: No
**BERT-VGG**: Yes
**BERT-NoAtt**: No

(d) *VizWiz_val_*00000698
Q: is this screen on the desktop or is there any word on here involving installing updates
**Gold**: No
**BERT-RG**: No
**BERT-ResNet**: Yes
**BERT-VGG**: Yes
**BERT-NoAtt**: No



(e) *VizWiz_val_*00002468
Q: Is my camera obscured?
**Gold**: Yes
**BERT-RG**: Yes
**BERT-ResNet**: Yes
**BERT-VGG**: No
**BERT-NoAtt**: No

(f) *VizWiz_val_*00003415
Q: Is this the shampoo?
**Gold**: Yes
**BERT-RG**: Yes
**BERT-ResNet**: No
**BERT-VGG**: No
**BERT-NoAtt**: No

Figure 5.8: Details of experimental examples to reveal our model's performance

technology to support the disabled community. Furthermore, the detailed examples in Section 5.5 prove the strength of our model and our components in both image and question's understanding. This success opens a novel prospect of considering the practical side of research to design and improve the more suitable and compact VQA models.

# Chapter 6

# General Visual Question Answering

## 6.1 Introduction

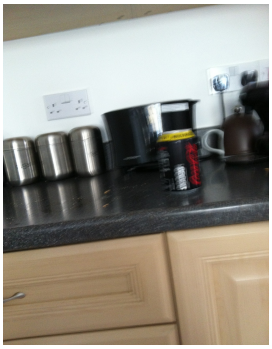
In the explosion of multi-media information, humans are living in tons of data in many categories. In most websites, it is easy enough to catch the appearance of texts, images, and videos. Obviously, in a multi-media real world, the complement of many kinds of data makes our lives more colorful and interesting. However, this incredible variety and complexity are the huge interest in research and application. The main question in multi-modal systems is how to analyze, benefit, and understand them to put them into practice. Generally, increasing human demand is the reason for research and development efforts in these areas. A new generation of multi-modal studies has emerged and has been growing rapidly in recent years.

At the beginning of multi-modal development, the combination between images and texts always gives the public interest in both research and application. It comes from the close relationship and appearance between them. In the usual way, texts are used to express and describe the content of images while images provide us more information about the context of texts. Despite the previous success in both texts and images, their cutting-edge studies are still new and have much potential. There are, however, still many existing challenges. Most of them come from the existing problems in texts and images. While a natural language is ambiguous and complicated, the visual structure is tightly

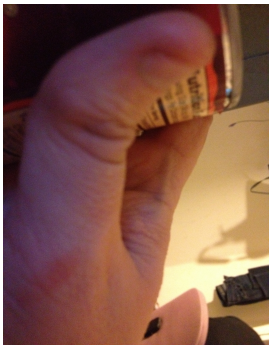
bound to the regional features. Therefore, to overcome these challenges, researchers should definitely take advantage of advances and technology in both Natural Language Processing (NLP) and Computer Vision (CV).

In the trend of visual-textual researches, there are many interesting and potential tasks such as Image Captioning [Li et al., 2020, Yun et al., 2019, Sharma et al., 2018], Vision Commonsense Reasoning [Wang et al., 2020c, Zellers et al., 2019], and Visual Question Answering [Su et al., 2020a, Yu et al., 2019, Goyal et al., 2019, Kazemi and Elqursh, 2017]. In these studies, the most important goal is to understand and combine the visual and textual features. Among the interesting tasks, our work pays attention to Visual Question Answering. Firstly, this task is quite balanced between text and image processing. The mainstream of previous VQA systems is in digesting the image and texts instead of considering the decoding phase later. Secondly, the independence between image and text understanding is promising to apply the strength in both two areas into our researches instead of building a huge architecture from scratch. However, our main subjective reason is in the practical benefit of Visual Question Answering. This task is much more interesting and useful in both research and application. Although the fundamental tasks are important and necessary, the application ones are also worth looking into deeply. Especially, the combination of basic and applied research is even more appreciated and encouraged.

Specifically, Visual Question Answering is to generate or determine the answer for a textual question about the content of an image. The examples of VQA samples are visualized in Figure 6.1. Obviously, the VQA system needs to understand the meaning of the image to point out the answer to its question. The answers belong to the quantity, quality, and detail of objects in images. Although detecting the content of image and object is familiar to Computer Vision, understanding and combining the content of image and question is quite novel and interesting in recent studies. In the complex structure of the textual question, it is too challenging to digest and extracts the important keywords which are related to the objects in the input image. Despite these existing challenges, VQA is indeed the fertile ground for multi-modal learning in images and texts. It motivates us to continue this challenging research.



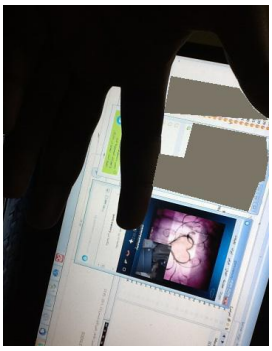
(a) *VizWiz_train_*00000001
**Q:** Can you tell me
what is in this can please?
**A:** {soda, coca cola (3), coke 0 (2),
unsuitable (2), coca cola 0, coke}
**Type:** Other



(b) *VizWiz_train_*00000011
**Q:** What is the sodium content
of this can of food?
**A:** {unanswerable (5), unsuitable (4),
insufficient photo quality}
**Type:** Unanswerable

(c) *VizWiz_val_*00023845
**Q:** Is this monitor on?
**A:** {yes (9), table (1)}

**Type:** Yes/No



(d) *VizWiz_train_*00023870
**Q:** How many fingers do I have?
**A:** {unanswerable (3), 10, 4 (2),
2 (2), 5, 3}
**Type:** Number

Figure 6.1: The examples of General Visual Question Answering in VizWiz-VQA 2020

Similar to most previous works, in our research, we consider general Visual Question Answering as the classification task. Although the number of the answer is large, this approach is much easier to access and deploy in practice than the generation tasks. The reason for this claim comes from the limited range of the VQA dataset. Instead of putting effort into finding the unlimited VQA system, most researches are geared towards solving pre-defined problems. Therefore, the coverage ratio of common answers is often quite large in VQA datasets [Kazemi and Elqursh, 2017]. This characteristic motivates us to regard this task as classification.

Traditionally, VQA systems often consist of three main components as feature extractor

and multi-modal fusion function. Specifically, the previous approaches often use the Convolution Neural Networks and Recurrent Neural Networks for image and question embedding [Goyal et al., 2019, Kazemi and Elqursh, 2017, Ren et al., 2015b]. However, it requires a huge architecture with a large number of trainable parameters to optimize from the scratch. The importance of the feature extractor is flouted due to the main objective of VQA.

Therefore, we propose to integrate the pre-trained models to simplify the role of this phase. Instead of building a huge architecture, the pre-trained models are indeed promising to take advantage of the other tasks into VQA. Among pre-trained models in NLP, the appearance of BERT has changed a lot in both research and application. With the strength of BERT in the language model, we integrate it into our question embedding to extract the textual features. On the other hand, the consistency between images and questions is extremely important to optimize effectively and easily. Therefore, we propose to integrate the novel Vision Transformer that is the most similar to BERT. Using the same Transformer architecture, Vision Transformer and BERT also get impressive performance in their tasks. Therefore, our integration is able to inherit the strength of these models into a brand-new domain, VQA.

The second component in Visual Question Answering is the fusion function between textual and visual features. In the traditional approaches, these features are often combined by a vector operation such as multiplication, addition, and so on. Recently, attention becomes more and more popular in both Computer Vision and Natural Language Processing. Especially, this mechanism has remarkable success in multi-modal tasks [Wang et al., 2020b, Huang et al., 2019, Anderson et al., 2018a]. This technique is also applied successfully in Visual Question Answering [Yu et al., 2019, Anderson et al., 2018a, Shih et al., 2016]. These works reveal that learning co-attention between images and texts is very important to capture their relationship and representation simultaneously. However, the attention is often considered in a single direction, which turn VQA systems into either image-aid or text-aid application instead of understanding both of them together. Therefore, inspired by the attention mechanism in Yu et al. [2019], we propose bi-directional

co-attention networks to learn and digest images and text simultaneously. Our attention allows the visual features to be supported by the question and vice versa. Besides, our attention uses the Transformer architecture as the main computation. Therefore, the consistency among the image, text, and attention brings us promising performance. In general, we propose a bi-directional co-attention on the Vision-Text Transformer model, BiCAtt VT-Transformer, for Visual Question Answering. This is the delegate and novel approach to take advantage of pre-trained models and Transformer architecture to understand and combine textual and visual features effectively in a bi-directional way. To prove the efficiency of our proposed method, we conduct all experiments in VizWiz-VQA 2020 dataset. Based on the experimental results and ablation studies, our BiCAtt VT-Transformer outperforms the existing baselines in the research and achieves promising results in real-world competition.

Our main contributions are presented as follows:

- By taking advantage of the strength of pre-trained models, we propose the simple and compact model to overcome the challenges in General Visual Question Answering. Instead of building a huge architecture for image and text extractor, our model integrates the recent successful architecture, Transformer, into Vison and Text. The delicate approach is effective and useful enough to capture the meaningful features in the VQA task.

- Inspired by the uni-directional co-attention networks, we propose a new kind of co-attention in a bi-directional mode. Through our attention, the content of image and text are considered simultaneously in the consistent architecture with the feature extractor.

- Through the experimental results and ablation studies, the strength of our proposed model is proved clearly in research. Especially, we also take part in the recent competition of VizWiz-VQA 2021 and archive the promising results at top-6. Even that, in some kinds of questions, our model outperforms the current SOTA systems in the leaderboard of this contest.

## 6.2 Related Works

In the era of multi-media, more and more data has risen rapidly every minute. It requires combining and integrate the advanced techniques of many areas to deal with the challenges in multi-modal tasks. At the beginning of the VQA problem, the previous approaches often utilize two dominant networks in image and text understanding. Specifically, the image is embedded by Convolution Neural Network while the question is learned by the Recurrent Neural Networks [Goyal et al., 2019, Ren et al., 2015a, Lu et al., 2015]. The strength of these approaches is the simplicity in deployment based on the well-known architecture in Computer Vision and Natural Language Processing. Despite the fundamental architecture, the combination between CNNs and RNNs with their variants is considered the important baseline in many VQA datasets and approaches.

In recent years, the success of pre-trained models in both CV and NLP spreads the new trend in these decades. The first impressive attempt in the VQA task is the work FT-VQA [Kazemi and Elqursh, 2017]. This model integrates the strength of the pre-trained image processing model to extract the visual features from the images. In question, it is traditionally utilized the RNNs variant, Long Short-term Memory Network (LSTM), to learn the sequence meaning. Specifically, FT-VQA uses the pre-trained ResNet-152 for image embedding. ResNet is improved from the CNNs architecture by the residual network to obtain and transmit the high-level features in images. It is pre-trained in the ImageNet dataset which is famous in the Image Classification task. By taking advantage of the pre-trained component in the image, FT-VQA obtains impressive performance in the VQA-v2 dataset and becomes the popular baseline in many previous approaches.

Early attempts of the VQA system also focus on the multi-modal fusion function to combine the visual and textual features. In the traditional approaches, this function is often vector operations such as point-wise multiplication, concatenation, and so on [Kolling et al., 2020]. Obviously, these operations are easy to implement and deploy in both research and application. The success of these VQA systems depends on the strength of feature extractors. However, in recent years, the appearance of attention inspired the

new generation of multi-modal fusion. The first attempt of learning visual attention from image regions via the question information is proposed in Word+Region Sel. model [Shih et al., 2016]. After the success of attention-based approaches, this kind of multi-modal fusion function becomes popular and important in most VQA systems [Kim et al., 2017, Anderson et al., 2018a].

In the flow of pre-trained models and regional attention in images, there are some approaches to integrate all of them into the VQA system such as FT-VQA [Kazemi and Elqursh, 2017], BERT-RG [Kazemi and Elqursh, 2017] with the stacked-attention mechanism. This attention is based on the textual features to intensify the regional images. Obviously, in this attention, the only image is enhanced by the attention scoring function. It eliminates the effect of the image into question. Through the attention in VQA, the simultaneous intensification in both textual and visual features is very essential. It is the reason for the introduction of the Co-Attention mechanism in many works such as BAN [Kim et al., 2018], DCN [Nguyen and Okatani, 2018], and so on.

The outstanding approach in Co-Attention works in VQA task is MCAN model [Yu et al., 2019]. This model integrates the powerful architecture of the Transformer to build the attention function. It emphasizes the importance of Co-Attention in VQA. Instead of digesting and scoring in one kind of modality (image and question), Co-Attention pays attention to both of them. Specifically, in the MCAN model, the image is embedded by Faster R-CNN [Ren et al., 2017] while the question is extracted in LSTM. In Attention phase, the visual and textual features are enhanced by the self-attention and guided-attention by the other. This attention inspired us about the importance of Co-Attention in VQA where it is necessary to extract and combine visual and textual features simultaneously. However, our attention is the improvement of MCAN with the simplicity in Co-Attention blocks. Besides, our model also considers the attention in bi-directional mechanism which is different from the uni-direction in MCAN work.

## 6.3 Methodology

In recent years, the strength of Transformer [Vaswani et al., 2017] architecture is proved in many tasks and systems [Su et al., 2020a],[Devlin et al., 2019], [Yu et al., 2019]. It is a powerful motivation for us to study and integrate Transformer into our tasks. However, the performance of this architecture is often accompanied by huge parameters. Therefore, starting from scratch is even less realistic. Recently, the trend of fine-tuning from the portable pre-trained models obtains more and more interest in both research and industry. In our proposed model, we also integrate the knowledge from pre-trained components into image and text processing. Through the goal of the compact and effective model, we propose to combine Vision and Text Transformer to extract the visual and textual features in a consistent process. The novelty of these components comes from our delicate consideration in Visual Question Answering. Besides, we also take advantage of Transformer architecture into our novel bi-direction co-attention network to combine and understand the relationship between visual and textual features. Our attention is enhanced the uni-directional co-attention networks from [Yu et al., 2019] into the novel and effective mechanism.

The content of this Chapter is organized as follows:

- Bi-directional Co-Attention Network (Section 6.3.1) provides the formulation and details of this mechanism to digest the visual and textual features simultaneously.
- Visual Question Answering Model (Section 6.3.2) shows the complete architecture of our model as combining all components in previous parts. Besides, we also mention the detail of the learning process.

### 6.3.1 Bi-directional Co-Attention Networks

In our architecture, we also take advantage of Question and Image Embedding in Section 3.2.1 and Section 3.2.2. specifically, questions are embedded by pre-trained BERT model while visual features are extracted by Vision Transformer with a stack of Average

Pooling 1D layer.

Comparing with the original VT-Transformer [Le et al., 2021b], our image feature extractor is connected by 1D-average pooling layers on intensified features of images in Equation 6.1. Obviously, with transfer-learning, our feature extractor inherits the strength of pre-trained BERT and Vision Transformer. Together with the efficiency of our feature extractor, our novel configuration is ideal to decrease the computation cost and deploy in the practice. The detail of our visual and textual feature extractors is presented in Figure 6.2.

$$z_k = \frac{1}{T}\sum_{i=s}^{s+T} z_i^L \tag{6.1}$$

where the stride of average pooling is $T$.

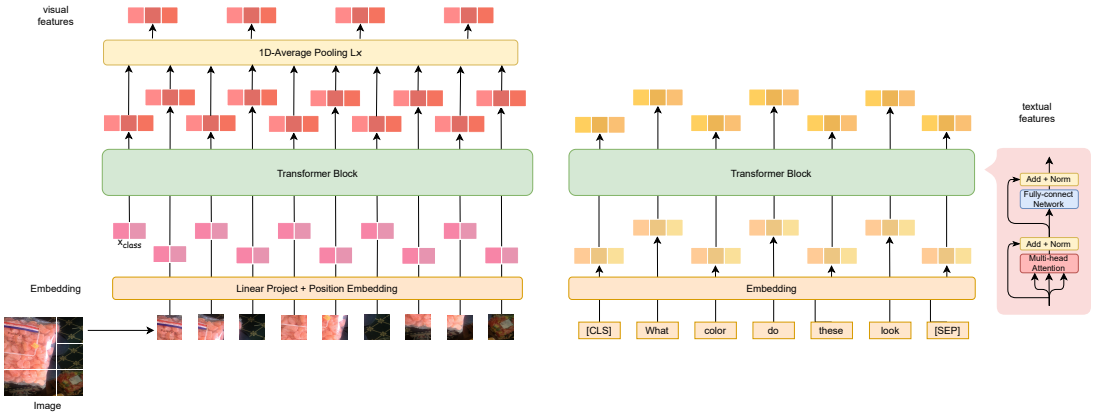



Figure 6.2: Feature Extractor utilizes BERT and Vision Transformer to understand the question and image via the strength of Transformer and transfer-learning mechanism

After extract the visual and textual features, the important question is how to combine them to understand and extract their relationship for predicting the answer. These features are often combined by a multi-modal fusion function such as multiplication, addition, and so on. The success of attention in both Computer Vision and Natural Language Processing is the great motivation to integrate it to learn the relationship in multi-modal systems.

Traditionally, the attention in fusion function is often connected to the global features of images and questions. Therefore, its effect is quite tiny to obtain the critical infor-

mation in the local regions. It means that the learning of images and question is done independently until meeting the fusion function. To overcome this problem, the Co-Attention model allows to learn the textual attention of question and visual attention of images simultaneously [Yu et al., 2018, Nam et al., 2017]. Inspired by the Deep Modular Co-Attention Networks [Yu et al., 2019], we propose bi-directional attention to learn the question and image from the other modality.

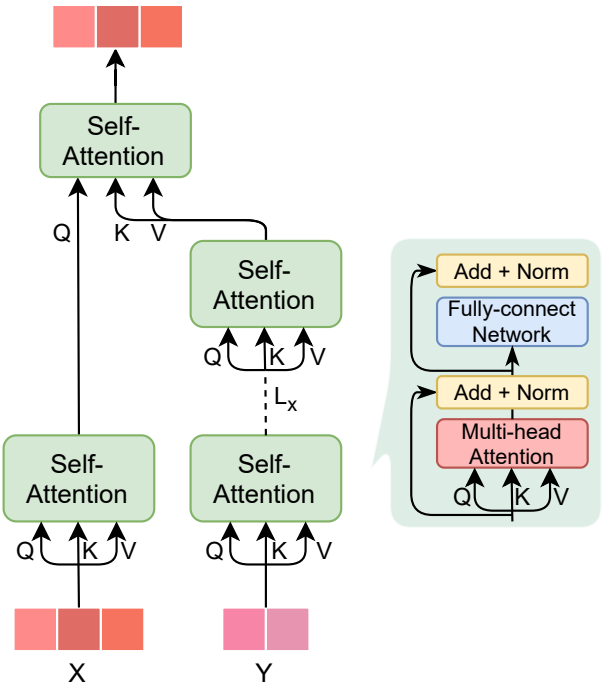



Figure 6.3: Deep Co-Attention Layer

The detail of our modified Deep Co-Attention layer is presented in Figure 6.3. Our layer is the direction attention layer, which means that the $CAtt$ of the image and the question is different from the $CAtt$ of the question and the image. Our attention consists of three Transformer blocks. Two lower blocks learn the self-attention of inputs to intensify the essential signal. The last block is to utilize the input $Y$ to guide the attention learning of $X$.

Mathematically, the self-attention of X and Y is calculated in Equation 6.2.

$$Z_X = Self - Attention(X, X, X); Z_Y = Self - Attention(Y, Y, Y) \tag{6.2}$$

The content of self-attention is similar to Equation 3.1 to Equation 3.11 where $q, K, V$ are the same. In the next block, the information of $Z_Y$ and $Z_X$ is utilized to calculate the attention weights to intensify the query $Z_X$ as the Equation 6.3.

$$\hat{X} = Self - Attention(Z_X, Z_Y, Z_Y) \tag{6.3}$$

It is the reason that we consider the $Z_X$ is guided by $Z_Y$. In general, our attention obtains two important points. Firstly, it is useful to obtain the locally essential feature in $X$ and $Y$ independently. After that, the information of $X$ is supported by the feature in $Y$. It means our attention is both local and global together.

In this way, the image, however, is only guided by the question, and vice versa. Obviously, only one component is intensified here while the VQA system needs meaningful features in both image and question. To deal with this problem, we propose to consider this attention in the bi-direction. It means that both image and question are guided together.

In our attention, we apply the Deep Co-Attention layer in both image and question. In this way, both local and global image and question are also learned by the other support. With the same architecture based on Transformer, the cost for attention decreases due to the consistency in image and question embedding, which is proved in the ablation studies. The strength of our attention is to combine the image and question simultaneously. Besides, the visual and textual understanding is intensified and supported by the other component instead of extracting features independently in the traditional approaches.

### 6.3.2 Visual Question Answering Model

After the bi-directional Co-Attention Network, the visual and textual features are augmented by the other together. The sequence of patch or token is flattened and normalized by the Attention Flatten layer into d-dimension space.

Then, in the final layer, we only use simple multiplication to combine the visual and textual features. Like traditional approaches, we also collect the most frequent answers as

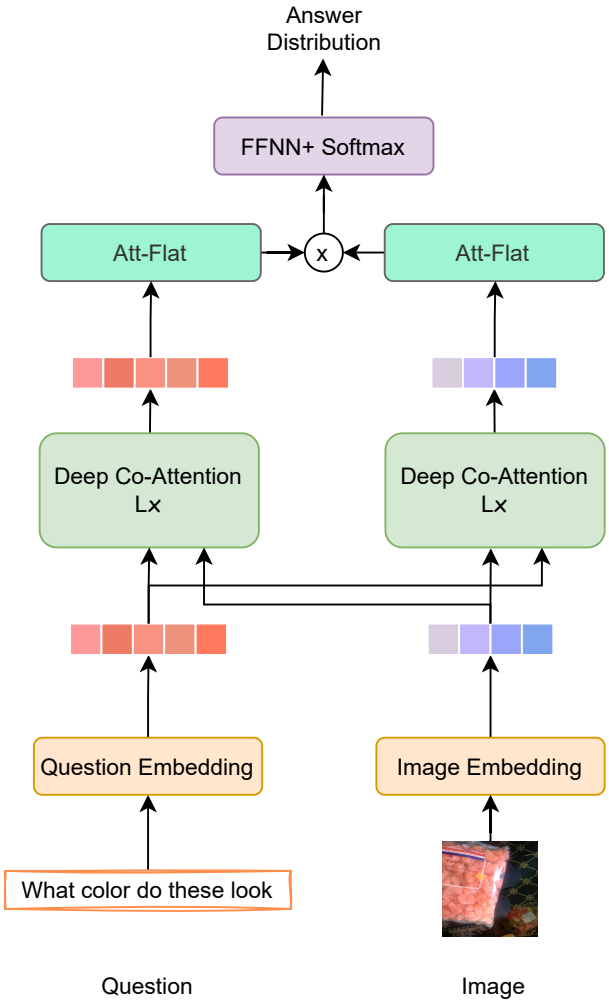



Figure 6.4: The detailed architecture of Bi-directional Co-Attention Visual Question Answering

the expected label. The classifier in our model is Multi-layer Perceptron with the softmax activation as follows.

$$y = Softmax\left(W_a^T(\hat{f}_I \otimes \hat{f}_Q) + b_a\right) \tag{6.4}$$

Where $W_a \in \mathbb{R}^{(d \times N)}$ and $b_a \in \mathbb{R}^{\mathbb{N}}$.

## 6.4 Experiment

### 6.4.1 Dataset & Evaluation Metric

In the goal of this research, we put our concentration on the VizWiz-VQA 2020 dataset [Gurari et al., 2018, 2019]. The detail of this dataset and its challenges is presented in Chapter 2. However, in the general Visual Question Answering, we apply some pre-processing steps to clean the input and gold answers.

- We analyze the frequency of answers and filter the most popular ones as the expected answer vocabulary.
- We eliminate the sample whose answer is not in the answer list.
- We eliminate the answer which does not appear in any sample.

The detail of the sample before and after pre-processing is presented in Table 6.1. The remaining rate in the train and validation set is approximately 96.2% and 92.9%. With the blind test set, the cover rate is challenging for our VQA system.

Table 6.1: The detail of dataset after pre-processing

| | Train | | Val | |
|---|---|---|---|---|
| | Before | After | Before | After |
| No. Sample | 20,523 | 19,744 | 4,319 | 4,014 |
| No. Answer | 41,299 | 2,953 | 10,950 | 2,953 |

In this task, we also utilize VQA Accuracy as the main evaluation metric. The detail of this measurement is presented in Chapter 5.4.2.

### 6.4.2 Experimental Settings

In our embedding, we take advantage of pre-trained models in Vision and Text Transformer to extract visual and textual features. Besides, we also propose the Bi-directional Co-Attention Networks to intensify the signal of question and image together. All details of our implementation are presented in Table 6.2. It is an important component to reproduce our model in the Pytorch language.

Table 6.2: The detail of experimental settings in our model

| Components | Model | Value |
|---|---|---|
| **Image Embedding** | **Vision Transformer** | B_16_imagenet1k |
| **Question Embedding** | **BERT** | bert-base-uncased |
| **Co-Attention** | **No. head** | 4 |
| | **Hidden size** | 1024 |
| | **Co-Attetion Layer** | 1 |
| **Dropout** | **Dropout** | 0.5 |
| **Training** | **Optimizer** | AdamW(lr = 5e-5, eps = 1e-8) |
| | **Loss** | CrossEntropy |

At the limit of our system, we only use the basic setting of both vision and text transformer. The explosion of computation cost increases rapidly in the Transformer blocks. However, these settings are ideal to deploy our compact and effective VQA system in our architecture.

### 6.4.3 Results

Table 6.3 presents the detail of our result and comparison. In our experiments, we compare our models with the following systems including:

- The powerful and popular CNNs:
  - CNN-LSTM [Goyal et al., 2019] uses the traditional CNNs layer for image embedding and LSTM for question embedding. This model does not use the strength of pre-trained models.
  - FT-VQA [Kazemi and Elqursh, 2017] uses the pre-trained ResNet for image embedding and LSTM for question embedding. Instead of CNN-LSTM, this model has appeared the pre-trained ResNet layer to extract visual features.
- BERT-RG is the modification of the original system whose classifier is changed into the MLP and softmax activation. It integrates the pre-trained BERT models for question embedding while pre-trained ResNet and VGG are utilized to extract the visual features.
- *Katya* is the winner of the VizWiz-VQA 2021 competition.

Obviously, our model obtains promising results against the existing approaches. It outperforms the famous systems in VQA. In the real-world competition, our model also achieves the top-7 in VizWiz-VQA 2021 competition. Our result is even quite equivalent to the winner *Katya* in some question classes.

Table 6.3: The detailed results of General Visual Question Answering in VizWiz-VQA 2020 dataset (Test-Standard)

| Model | Yes/No | Number | Other | Unanswerable | Overall |
|---|---|---|---|---|---|
| **CNN-LSTM** Goyal et al. [2019] | 54.06 | 17.07 | 25.09 | 77.97 | 40.42 |
| **FT-VQA** Kazemi and Elqursh [2017] | 58.17 | 20.6 | 18.21 | 76.56 | 35.65 |
| **BERT-RG** Le et al. [2020] | 71.82 | 15.72 | 20.52 | 85.44 | 40.11 |
| **VT-Transformer** Le et al. [2021b] | 73.03 | 12.17 | 25.15 | 69.95 | 37.24 |
| **LXMERT** Tan and Bansal [2019a] | 66.48 | 17.99 | 20.47 | 65.22 | 34.85 |
| **OL-LXMERT** | 70.04 | 16.40 | 27.75 | 72.57 | 41.85 |
| **Our model BiCAtt VT-Transformer** | **73.64** | **24.12** | **34.86** | **82.61** | **49.19** |
| **Katya 2021-Rank#1** | **80.52** | **27.37** | **40.92** | **86.82** | **54.76** |
| **Our model 2021-Rank#7** | 73.64 | 24.12 | 34.86 | 82.61 | 49.19 |

Firstly, our model is more potential than the CNNs approaches. The performance of our model is better than CNN-LSTM [Goyal et al., 2019] and ResNet-LSTM [Kazemi and Elqursh, 2017]. This success comes from the strength of pre-trained Vision and Text Transformer. Even that, the comparison between our model and BERT-RG [Le et al., 2020] proves that Transformer architecture is more potential than CNNs-based models. Although our model can not achieve the winner in VizWiz-VQA 2021 competition, it is promising and approximate to *Kayta* in Number and Unanswerable. Besides, the potential of our system is very large due to the volumes of pre-trained components in our model.

### 6.4.4 Ablation Studies

To prove the novelty and effect of our proposed components, we would like to present the detail of ablation studies from some viewpoints. The first one is about the strength of Vision Transformer Dosovitskiy et al. [2021] against the CNNs and its variant. In this part, we compare our Vision Transformer against pre-trained VGG-16 models. VGG [Simonyan and Zisserman, 2014] is the famous variant of CNNs model in image understanding. It has 16 deep CNN layers with the fixed-size kernel to improve the vanilla AlexNet [Krizhevsky et al., 2012]. Obviously, the Vision Transformer proves its strength in all question types.

Table 6.4: The strength of Vision Transformer against VGG (Test-dev)

| Image Embedding | Yes/No | Number | Other | Unanswerable | Overall |
|---|---|---|---|---|---|
| **VGG** | 71.91 | 20.11 | 32.44 | 81.19 | 47.48 |
| **Vision Transformer** | **76.59** | **28.04** | **36.47** | **81.04** | **50.45** |

The second component of our ablation studies is the effect of the number of Attention Layer. In the limit of our computation resource, we present the result of our comparison until 4 layers. The number of layer $L$ here implies the first Transformer block in the Deep Co-Attention Layer of Figure 6.3. It means the input X is intensified in L self-attention layers before being put into the guided self-attention block. The detail of our comparison is presented in Table 6.5. The discovery here is the decrease of performance through the increase of the attention layer. In our observation, it comes from the consistency of the feature extractor and our attention. In the image and question embedding, the visual and textual are emphasized by the self-attention function. Therefore, the increasing of this attention in the modality only brings us the explosion of computation cost.

Table 6.5: The effect of BiCAtt Layer into our attention (Test-dev)

| Co-Attention Layer | Yes/No | Number | Other | Unanswerable | Overall |
|---|---|---|---|---|---|
| **1** | **76.59** | **28.04** | **36.47** | **81.04** | **50.45** |
| **2** | 74.53 | 17.46 | 35.03 | 79.93 | 48.92 |
| **3** | 73.97 | 20.63 | 34.84 | 79.84 | 48.80 |
| **4** | 69.85 | 21.16 | 34.59 | 75.56 | 47.27 |

The last part of our comparison is the direction of the Deep Co-Attention Layer. In our proposed model, the bi-direction plays an important role to enhance the performance of our completed system. Specifically, we compare the uni-direction between image and question against the bi-directional mechanism. The notation $Img \rightarrow Qst$ implies the question guided by the image. It means the input X is an question and Y is a image.

Table 6.6: The comparison between uni-direction and bi-direction Co-Attention Network in VizWiz-VQA 2020 (Test-dev)

| Model | Yes/No | Number | Other | Unanswerable | Overall |
|---|---|---|---|---|---|
| **Img → Qst** | 71.54 | 24.34 | 34.88 | 80.05 | 48.83 |
| **Qst → Img** | 69.85 | 17.99 | 34.02 | 83.42 | 49.02 |
| **Our model** | **76.59** | **28.04** | **36.47** | **81.04** | **50.45** |

Through the detailed comparison in Table 6.6, the bi-direction has an impressive effect on our model's success. In both uni-direction, our proposed direction proves its strength and performance in all question types.

## 6.5 Hierarchical VQA Framework

Together with the successes of four components in our works, the main goal of our research is to overcome the typical challenges of VQA for blind people via a hierarchical framework. The detail of our proposed architecture is presented in Section 1.3. In our system, the VQA sample is filtered by the answerability threshold. In our completed system, if the answerability score of the sample is higher than a pre-defined threshold, it is considered as an answerable sample. Next, valid samples are divided into two groups by Visual Question Classification. As we mentioned in Chapter 5 and Chapter 6, our consideration is highly suitable for this typical domain to narrow down the searching space of answer prediction. With our configuration, each specific type of question is optimized independently. Its proper domain and loss function is ideal to enhance the system's performance.

Obviously, in our architecture, the answerability threshold plays an important role in determining the early answer for the VQA sample. Particularly, in our completed system, invalid samples are assigned to the "unanswerable" class in the answer space. In

Visual Question Classification, the samples are divided into Other (including Number, Unanswerable, and Other in VizWiz-VQA configuration) and Yes/No questions. The decision is based on the dominant class via the output of the softmax layer in Figure 4.4. Finally, we utilize the Yes/No VQA and General VQA systems to reveal the corresponding answers of samples.

However, the VizWiz-VQA dataset is a private evaluation without any information in the test set. Unfortunately, it requires the gold label of Answerability Prediction to evaluate our system performance. It comes from the diversity of answers in the annotation. Although a sample is considered unanswerability, its answer may not be "unanswerable". There are a lot of options in this case such as "unsuitable", "can not answer", "uncountable", and so on. Even though we can determine the answerability of the sample, we can not find out a suitable answer for the private evaluation of the VizWiz-VQA dataset. With the above challenges, we recommend utilizing the same configuration of Visual Question Classification to evaluate our completed systems. It means that the validation set of the original VizWiz-VQA is used as the test set for our evaluation. The detail of our modified dataset is similar to the VQC task and is presented in Section 4.4.1. We also optimize four sub-models with the same configuration as the VQC task.

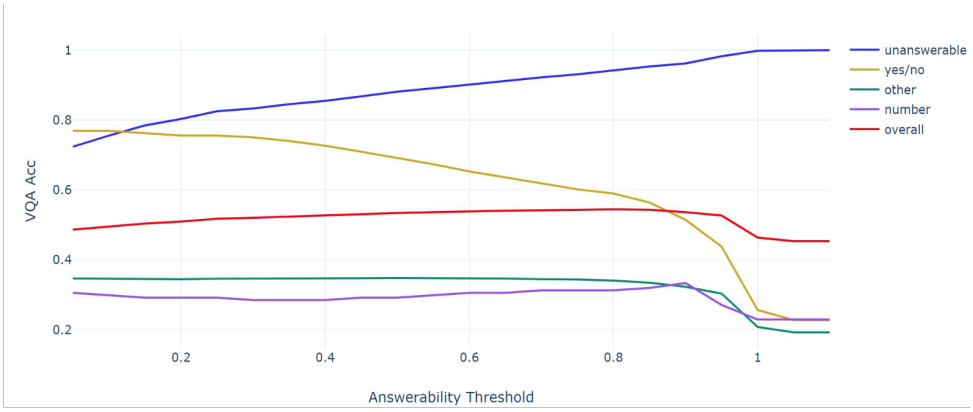



Figure 6.5: The effect of Answerability Threshold in modified VizWiz-VQA 2020

Following the flow of our hierarchical framework, we would like to reveal the strength of our proposed systems against the full-stack VQA systems, BiCAtt VT-Transformer. To prove the performance of our hierarchical architecture, we recommend three main

comparisons as follows:

- The effect of Answerability Threshold in our Hierarchical framework
- The comparison between our system and BiCAtt VT-Transformer via Answerability Threshold
- The strength of our proposed system against BiCAtt VT-Transformer without Answerability Threshold

Firstly, as we mentioned above, the Answerability Threshold is used to filter the invalid samples for VQA systems. With the change of this threshold, it is easy to control the percentage of valid samples before extracting the correct answers. In Figure 6.5, when the threshold is increased, the performance of the system is also enhanced overall.

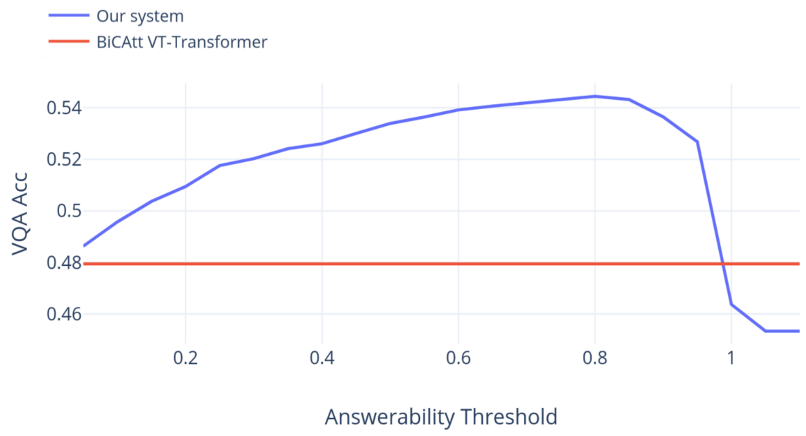



Figure 6.6: Comparison between our hierarchical architecture and BiCAtt VT-Transformer via four types of question in modified VizWiz-VQA

However, from the limit of 0.9, the system works worse and worse. The reason for this phenomenon comes from the distribution of unanswerable samples in the VizWiz-VQA dataset. As it is shown in Section 4.4.1, there are a lot of unanswerable samples in this dataset. Therefore, it is obviously challenging for the previous approaches. In our systems, we allow users to define their threshold to adapt it into their systems. This

configuration is ideal to deploy the VQA systems into practice, especially in the domain of blind people.

Together with the flexibility of the Answerability threshold, the second factor we would like to emphasize is the strength of our proposed framework against BiCAtt VT-Transformer, a general model of VQA for blind people. The performance of our proposed framework is presented in Figure 6.6. The trend of this comparison is similar to the previous one. When the filtering condition is tightened, the strength of our completed system increases quickly until it gets the limit. It proves that Answerability Prediction is highly important in the VQA systems especially for blind people. It is also a typical characteristic of VizWiz-VQA against the other VQA datasets and domains.

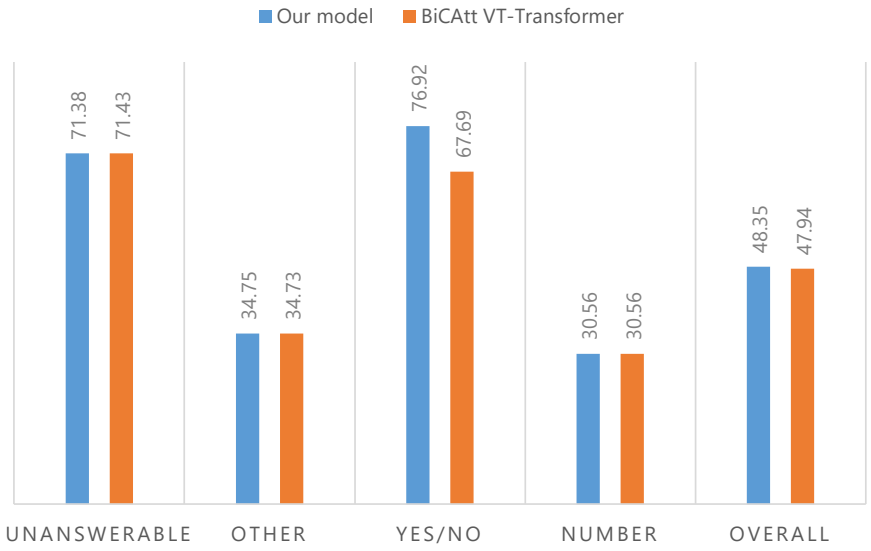



Figure 6.7: The strength of hierarchical framework against BiCAtt VT-Transformer without Answerability Prediction

Another difference between our framework and the previous ones is on Visual Question Classification. To prove the strength of this component, we eliminate the answerability filter in this comparison. Without the constraint of filter, our proposed framework also obtains the success of the VQA task against the general architecture, BiCAtt VT-Transformer. It reveals that optimization in the specific question type plays an important role in VQA for blind people. It is useful to narrow down the answer's space and orient the decision of VQA systems. In general, via three comparisons, our hierarchical architecture

achieves promising and competitive performance against the previous VQA systems.

## 6.6 Summary

In this work, we propose the BiCAtt-VT Transformer model to integrate the strength of the Transformer in both feature extractor and multi-modal fusion function. Our approach integrates the Vision and Text Transformer to understand the image and question the consistency of Transformer architecture. By taking advantage of pre-trained models, our model is more compact and effective than the CNNs and RNNs approaches.

Inspired by the MCAN attention, we propose the bi-direction Co-Attention Network to combine the visual and textual features. This attention allows one to learn these features simultaneously through the support of the other. In our attention, the visual and textual features are intensified together in the Transformer block. With the consistency of the Transformer in feature extractors and attention, our model decreases the computation cost of the attention layer. It obtains impressive performance in only one attention layer.

In experiments and ablation studies, the comparison between our model and existing baselines proves the effectiveness and novelty of our models. Even though our model also obtains promising performance in real-world competition. The success of our model comes from the strength of our proposed architecture and bi-directional attention.

# Chapter 7

# Conclusions and Future Work

## 7.1 Conclusions

Our thesis is motivated by the fact that Visual Question Answering on a multi-modal approach will benefit many applications and research, especially on supporting blind people to overcome their difficulties. Besides, the success of pre-trained models and Deep Learning are promising approaches for solving that task. Together with the scientific goals, our work is aimed at raising the interest and concern of our community to help the disabled, especially the blind by intelligent approaches and systems.

The main contributions of this dissertation are summarized as follows:

- **Answerability Prediction** (Chapter 3): Through our observation, we introduce the novel problem statement in the brand-new topic in VQA. We propose the new VT-Transformer to predict the answerability score of VQA samples. Our model is the combination of two powerful components including Vision Transformer for image process and BERT (Text Transformer) for question understanding. With the strength of our model, it obtains impressive performance in the VizWiz-VQA 2020 and 2021. The success of our model proves the novelty of our approach and proposal to this problem as well as the strength of our VT-Transformer in the real-world dataset.

- **Visual Question Classification** (Chapter 4): We propose the novel task called Visual Question Classification to take advantage of hidden information in most VQA datasets. Furthermore, we also point out the bottle-neck of current vision-language models through depending on external Object Detection models. To overcome this limit, we propose a Object-less LXMERT models to make use of pre-trained LXMERT [Tan and Bansal [2019a]] system with our virtual object generator. By eliminating the role of Object Detection, our model obtains the promising results against the competitive baselines.

- **Yes/No Visual Question Answering** (Chapter 5): We reveal the important role of Yes/No question in both research and application. In our proposed model, we integrate the strength of pre-trained models to extract the visual and textual features. In addition, we also propose the mechanism to combine the residual and global features in images to enhance the visual features. The stacked attention in our model also brings the effective combination between regional features in the image and textual features in question. Through the experimental comparison, our Yes/No VQA model obtains a strong performance in both research and real-world competition of VizWiz-VQA 2020.

- **General Question Answering** (Chap 6): The proposed BiCAtt VT-Transformer architecture, which employs multiple pre-trained word models and bi-direction Co-Attention Network, consistently obtains competitive performances on the VizWiz-VQA dataset. The advantage is that the model allows the image and question understanding to be done simultaneously in their interaction. Our bi-direction Co-Attention is to take advantage of the image to intensify the textual features and vice versa. Compared to the existing methods in the VizWiz-VQA dataset and competition, our model obtains competitive and promising results in the fundamental setting of pre-trained Transformer models.

## 7.2 Future Work

In the future, our next study will focus on the following things:

- **Answerability Prediction** (Chapter 3): For the next future work, we intend to investigate the new multi-modal fusion functions to combine the visual and textual features. Despite the powerful feature extractor, the attention and fusion function is absolutely potential to enhance our model. Besides, we also study the unanswerability characteristic in many datasets in VQA as well as in many multi-modal tasks.

- **Visual Question Classification** (Chapter 4): With our observation as well as the trend in vision-language models, we intend to combine our virtual object generator and Object Detection models to deal with all kinds of images. Despite the promising result of our models, the object-based features are also the fertile features to enhance in object-less domain.

- **Yes/No Visual Question Answering** (Chapter 5): In our work, we just researched on the VizWiz-VQA dataset to prove the performance of our model. It is potential and necessary to put more effort into many kinds of datasets such CLEVER [Johnson et al. [2017a]], VQA v2.0 [Goyal et al. [2017]]. The new experiment is important to help us consolidate the importance of Yes/No questions in the research and application.

- **General Question Answering** (Chapter 6): Through the competitive results in VizWiz-VQA 2021 competition, we realize that our model needs to improve lightly to deploy in practice. Therefore, we would like to integrate the bi-direction mechanism in the image and question understanding phase. It is obviously necessary to learn and understand the visual and textual features together instead of through the Co-Attention for the improvement.

- **Explainable Visual Question Answering**: Through the studies on VQA, we intend to analyze and extract the relationship between objects in questions and

images to build the explainable structure in the VQA sample. Even that it allows to incorporate the external knowledge graph into the visual knowledge graph. This structure helps us to understand, modify and verify the decision for the practical application. Besides, this system is expected to integrate the question classification module to limit the answer vocabulary which is also the existing problem in VQA in the application.

# Bibliography

P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018a.

P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018b.

S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: visual question answering. *CoRR*, abs/1505.00468, 2015. URL http://arxiv.org/abs/1505.00468.

N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, editors, *Computer Vision – ECCV 2020*, pages 213–229, Cham, 2020. Springer International Publishing.

Y.-C. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu. Uniter: Universal image-text representation learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, August 2020.

Y.-S. Chuang, C.-L. Liu, H. yi Lee, and L. shan Lee. SpeechBERT: An Audio-and-Text

Jointly Learned Language Model for End-to-End Spoken Question Answering. In *Proc. Interspeech 2020*, pages 4168–4172, 2020. doi: 10.21437/Interspeech.2020-1570.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Y. Goyal, T. Khot, A. Agrawal, D. Summers-Stay, D. Batra, and D. Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *Int. J. Comput. Vision*, 127(4):398–414, Apr. 2019. ISSN 0920-5691. doi: 10.1007/s11263-018-1116-0.

D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

D. Gurari, Q. Li, C. Lin, Y. Zhao, A. Guo, A. Stangl, and J. P. Bigham. Vizwiz-priv: A dataset for recognizing the presence and purpose of private visual information in images taken by blind people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8): 1735–1780, Nov. 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

L. Huang, W. Wang, J. Chen, and X.-Y. Wei. Attention on attention for image captioning. In *International Conference on Computer Vision*, 2019.

D. A. Hudson and C. D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017a.

J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017b.

V. Kazemi and A. Elqursh. Show, ask, attend, and answer: A strong baseline for visual question answering. *CoRR*, abs/1704.03162, 2017.

D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 2611–2624. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1b84c4cee2b8b3d823b30e2d604b1878-Paper.pdf.

J. Kim, K. W. On, W. Lim, J. Kim, J. Ha, and B. Zhang. Hadamard product for low-rank bilinear pooling. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

J.-H. Kim, J. Jun, and B.-T. Zhang. Bilinear Attention Networks. In *Advances in Neural Information Processing Systems 31*, pages 1571–1581, 2018.

Y. Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, Oct. 2014a. Association for Computational Linguistics. doi: 10.3115/v1/D14-1181. URL https://www.aclweb.org/anthology/D14-1181.

Y. Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, Oct. 2014b. doi: 10.3115/v1/D14-1181.

C. Kolling, J. Wehrmann, and R. C. Barros. Component analysis for visual question answering architectures. In *2020 International Joint Conference on Neural Networks, IJCNN 2020, Glasgow, United Kingdom, July 19-24, 2020*, pages 1–8. IEEE, 2020.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

T. Le, N. Tien Huy, and N. Le Minh. Integrating transformer into global and residual image feature extractor in visual question answering for blind people. In *2020 12th International Conference on Knowledge and Systems Engineering (KSE)*, pages 31–36, 2020. doi: 10.1109/KSE50997.2020.9287539.

T. Le, H. T. Nguyen, and M. L. Nguyen. Multi visual and textual embedding on visual question answering for blind people. *Neurocomputing*, 2021a. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2021.08.117.

T. Le, H. T. Nguyen, and M. L. Nguyen. Vision and text transformer for predicting answerability on visual question answering. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 934–938, 2021b. doi: 10.1109/ICIP42928.2021.9506796.

X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, Y. Choi, and J. Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, August 2020.

T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

Y. Lin, Y. Meng, X. Sun, Q. Han, K. Kuang, J. Li, and F. Wu. Bertgcn: Transductive text classification by combining gnn and bert. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, Aug. 2021.

J. Lu, X. Lin, D. Batra, and D. Parikh. Deeper lstm and normalized cnn visual question answering model. https://github.com/VT-vision-lab/VQA_LSTM_CNN, 2015.

J. Lu, V. Goswami, M. Rohrbach, D. Parikh, and S. Lee. 12-in-1: Multi-task vision and language representation learning. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In Y. Bengio and Y. LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.

H. Nam, J.-W. Ha, and J. Kim. Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

D.-K. Nguyen and T. Okatani. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran. Image transformer. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4055–4064. PMLR, 10–15 Jul 2018. URL http://proceedings.mlr.press/v80/parmar18a.html.

J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

M. Ren, R. Kiros, and R. S. Zemel. Exploring models and data for image question answering. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, page 2953–2961, Cambridge, MA, USA, 2015a. MIT Press.

M. Ren, R. Kiros, and R. S. Zemel. Exploring models and data for image question answering. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, page 2953–2961, Cambridge, MA, USA, 2015b. MIT Press.

S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6): 1137–1149, June 2017. ISSN 0162-8828. doi: 10.1109/TPAMI.2016.2577031.

P. Sharma, N. Ding, S. Goodman, and R. Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1238.

Y. Shi, T. Furlanello, S. Zha, and A. Anandkumar. Question type guided attention in visual question answering. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

K. J. Shih, S. Singh, and D. Hoiem. Where to look: Focus regions for visual question answering. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4613–4621, 2016. doi: 10.1109/CVPR.2016.499.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2020a.

W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2020b.

H. Tan and M. Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language*

*Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China, Nov. 2019a. doi: 10.18653/v1/D19-1514.

H. Tan and M. Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China, Nov. 2019b. Association for Computational Linguistics. doi: 10.18653/v1/D19-1514. URL https://www.aclweb.org/anthology/D19-1514.

M. Tan and Q. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 09–15 Jun 2019.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017.

H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, and L.-C. Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, editors, *Computer Vision – ECCV 2020*, pages 108–126, Cham, 2020a. Springer International Publishing.

J. Wang, J. Tang, and J. Luo. Multimodal attention with image text spatial relationship for ocr-based image captioning. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 4337–4345, New York, NY, USA, 2020b. Association for Computing Machinery. ISBN 9781450379885. doi: 10.1145/3394171.3413753. URL https://doi.org/10.1145/3394171.3413753.

T. Wang, J. Huang, H. Zhang, and Q. Sun. Visual commonsense r-cnn. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020c.

Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21–29, 2016.

Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

L. Yao, C. Mao, and Y. Luo. Graph convolutional networks for text classification. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7370–7377, 2019. doi: 10.1609/aaai.v33i01.33017370.

Z. Yu, J. Yu, C. Xiang, J. Fan, and D. Tao. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE Transactions on Neural Networks and Learning Systems*, 29(12):5947–5959, 2018.

Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

# Publications and Awards

## Journals

[1] Tung Le, Huy Tien Nguyen, and Minh Le Nguyen. "Multi Visual and Textual Embedding on Visual Question Answering for Blind People", Volume 465, Pages 451-464, Neurocomputing, Sep. 2021.

## Conference papers

[2] Tung Le, Khoa Pho, Thong Bui, Huy Tien Nguyen, and Minh Le Nguyen. "Object-less Vision-language Model on Visual Question Classification for Blind People" 14th International Conference on Agents and Artificial Intelligence (ICAART 2022), Feb. 2022.

[3] Tung Le, Thong Bui, Huy Tien Nguyen, and Minh Le Nguyen. "Bi-direction Co-Attention Network on Visual Question Answering for Blind People" 14th International Conference on Machine Vision (ICMV 2021), Nov. 2021.

[4] Tung Le, Huy Tien Nguyen, and Minh Le Nguyen. "Vision and Text Transformer for Predicting Answerability on Visual Question Answering", 2021 IEEE International Conference on Image Processing (ICIP), pp. 934-938, Sep. 2021.

[5] Tung Le, Nguyen Tien Huy, and Nguyen Le Minh. "Integrating Transformer into Global and Residual Image Feature Extractor in Visual Question Answering for

Blind People." Knowledge and System Engineering (KSE) 12th International Conference on IEEE, pp 31-36, Nov. 2020.

[6] Tung Le and Nguyen Le Minh, "Integration of Textual Discriminator into Multi-head Attention Model in Relation Extraction", Information system WINTER FESTA Episode 5, 2019.

[7] Nguyen Tien Huy, Le Thanh Tung, and Nguyen Le Minh. "Opinions Summarization: Aspect Similarity Recognition Relaxes The Constraint of Predefined Aspects." Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), pp 487-496, Sep. 2019.

[8] Tung Le and Nguyen Le Minh, "Combined Objective Function in Deep Learning Model for Abstractive Summarization", Proceedings of the Ninth International Symposium on Information and Communication Technology (SOICT 2018), pp 84-91, Dec. 2018.