VIETNAM GENERAL CONFEDERATION OF LABOR

TON DUC THANG UNIVERSITY

FACULTY OF INFORMATION TECHNOLOGY

PHAM HUYNH TIN – 522H0150

NGUYEN TRUNG THANG – 522H0145

MIDTERM REPORT

DEEP LEARNING

HO CHI MINH CITY, 2025

VIETNAM GENERAL CONFEDERATION OF LABOR

TON DUC THANG UNIVERSITY

FACULTY OF INFORMATION TECHNOLOGY

PHAM HUYNH TIN – 522H0150

NGUYEN TRUNG THANG – 522H0145

MIDTERM REPORT

DEEP LEARNING

Instructor

Assoc. PhD. Le Anh Cuong

HO CHI MINH CITY, 2025

ACKNOWLEDGEMENT

Dear Mr. Le Anh Cuong,

We would like to express our sincere and profound thanks to the teacher for the dedication and valuable sharing during the learning process. The enthusiasm, erudite knowledge as well as the encouragement of the teacher helped us not only master the knowledge but also inspired to explore more deeply and be more passionate about the field of study. The patience and dedication of the teacher have opened us with many knowledge doors, helping us to confidently move on the path of conquering new challenges.

We are extremely grateful for the precious moments that he has given us, advice and priceless motivation. Once again, sincerely thank you for all.

Sincerely welcome!

Ho Chi Minh City, March 24th, 2025

Authors

(Sign and write full name)

THE WORK IS COMPLETED

AT TON DUC THANG UNIVERSITY

I hereby declare that this is my own research project and is under the scientific guidance of Assoc. PhD. Le Anh Cuong. The research content and results in this topic are honest and have not been published in any form before. The data in the tables for analysis, comments, and evaluation were collected by the author from different sources and clearly stated in the reference section.

Project also uses a number of comments, assessments as well as data from other authors and other organizations, all with citations and source notes.

If any fraud is detected, I will take full responsibility for the content of my Project. Ton Duc Thang University is not involved in copyright violations caused by me during the implementation process (if any).

Ho Chi Minh City, March 24th, 2025

Authors

(Sign and write full name)

ABSTRACT

Visual Question Answering (VQA) is a challenging task that requires an AI model to understand both visual and textual information to generate accurate responses. In this project, we focus on developing a VQA system specifically designed for fruit and vegetable recognition. Our approach integrates deep learning techniques, combining a ResNet18 convolutional neural network (CNN) for image feature extraction and a Long Short-Term Memory (LSTM) network for question encoding. The Bootstrapped Language-Image Pretraining (BLIP) model is leveraged for automatic question-answer pair generation, enhancing data diversity and reducing manual labeling effort.

The dataset used in this study, sourced from Kaggle's Fruit and Vegetable Image Recognition dataset, consists of over 33,000 images across 36 categories. Extensive preprocessing, including image augmentation and tokenization, was performed to optimize training efficiency. The model was trained using the Adam optimizer, with Automatic Mixed Precision (AMP) and Gradient Accumulation techniques applied to improve computational efficiency.

Experimental results demonstrate that our VQA system achieves high accuracy in answering domain-specific questions about fruits and vegetables. The system performs well in classification and description tasks but faces challenges with complex reasoning-based queries. Future work aims to expand the dataset, incorporate transformer-based architectures, and integrate external knowledge sources to enhance model performance.

This research highlights the potential applications of domain-specific VQA in automated retail, dietary assistance, agricultural monitoring, and accessibility for visually impaired individuals, paving the way for real-world implementations.

TABLE OF CONTENTS

ABBREVIATIONS **6**

CHAPTER 1. OVERVIEW **1**

1.1 Introduction 1

1.2 Motivation and Applications 1

1.3 Problem Statement 2

CHAPTER 2. LITERATURE REVIEW **4**

2.1 VQA and Domain Specialization 4

2.2 Image Feature Extraction with ResNet18 4

2.3 Question Encoding with BertTokenizer and LSTM 5

2.4 Data Generation with BLIP 5

2.5 Training Optimization: AMP and Gradient Accumulation 6

CHAPTER 3. METHODOLOGY **7**

3.1 Dataset and Preprocessing 7

1.1 Dataset description *7*

1.2 Question - Answer Pair Generation *7*

1.3 Data Preprocessing *8*

3.2 Model Architecture 8

1.1 Overview *8*

1.2 Image Feature Extraction *8*

1.3 Question Encoding *9*

1.4 Multimodal Fusion and Answer prediction *9*

3.3 Training Procedure and Optimization 9

CHAPTER 4. EXPERIMENTS AND RESULTS 11

4.1 Results and Analysis 11

4.2 Discussion 11

CHAPTER 5. CONCLUSION 12

5.1 Summary and Limitations 12

5.2 Future Work 13

5.3 Concluding Remarks 14

REFERENCES 15

ABBREVIATIONS

VQA	Visual Question Answering
NLP	Natural Language Processing
AI	Artificial Intelligence
X-rays	X-radiation
LSTM	Long Short-Term Memory
CNNs	Convolutional Neural Networks
RNN	Recurrent Neural Network
BLIP	Bootstrapped Language-Image Pretraining
AMP	Automatic Mixed Precision
GPU	Graphics Processing Unit
ReLU	Rectified Linear Unit
SGD	Stochastic Gradient Descent
LLMs	Large Language Models
RLHF	Reinforcement Learning from Human Feedback
MRI	Magnetic Resonance Imaging

CHAPTER 1. OVERVIEW

1.1 Introduction

Visual Question Answering (VQA) has emerged as a prominent area of research at the intersection of computer vision and natural language processing (NLP). The core challenge lies in developing systems that can comprehend both the visual content of an image and the semantic meaning of a natural language question, and subsequently generate an accurate and contextually relevant answer. The complexities involved in VQA have propelled its importance as a benchmark problem for evaluating AI systems' ability to integrate multimodal information, mimicking aspects of human-level understanding.

This project explores VQA within a constrained, yet practically relevant domain: the identification and classification of fruits and vegetables. By narrowing the scope to this specific category of images, we aim to create a VQA system with high accuracy and efficiency, while addressing the challenges of data sparsity and domain-specific knowledge. This targeted approach allows us to investigate and optimize specific techniques for feature extraction, question encoding, and multimodal fusion.

1.2 Motivation and Applications

The motivation behind focusing on a fruit and vegetable VQA system stems from its potential applications across various sectors:

Automated Retail: Enabling automated checkout systems in supermarkets by verifying produce items based on visual input and user questions ("Is this an organic apple?").

Dietary Assistance: Developing mobile applications that allow users to identify food items and obtain nutritional information by simply taking a picture and asking questions ("How many calories are in this banana?").

Agricultural Monitoring: Supporting smart farming initiatives by assisting farmers in diagnosing crop diseases or identifying produce quality based on visual inspection ("Are these tomatoes ripe enough for harvest?").

Educational Tools: Creating interactive learning applications for children to learn about different types of fruits and vegetables ("What is the name of this red fruit?").

Accessibility: Providing visual assistance to individuals with visual impairments, enabling them to identify and learn about food items. These diverse applications highlight the value of a specialized VQA system that can reliably and accurately answer questions about fruit and vegetable imagery.

1.3 Problem Statement

The core problem addressed in this project is the development of a deep learning-based system capable of answering natural language questions about images of fruits and vegetables. Given an input image I belonging to the domain of fruits and vegetables and a natural language question Q, the system must generate an answer A that is semantically consistent with both I and Q.

This problem entails several key challenges:

Image Understanding: Accurately extracting relevant visual features from the input image, capturing information about object shape, color, texture, and context.

Language Understanding: Parsing and understanding the semantic meaning of the input question, including identifying key entities, relationships, and constraints.

Multimodal Fusion: Effectively integrating visual and textual information to reason about the relationship between the image and the question.

Knowledge Representation: Representing domain-specific knowledge about fruits and vegetables, such as their categories, attributes, and relationships.

To address these challenges, our project focuses on developing a system that can answer questions of the following types:

Categorization: "Is this a fruit or a vegetable?", "Is this a citrus fruit?"

Identification: "What fruit is in the image?", "What kind of vegetable is this?"

Description: "Describe the image." (Generating a caption)

CHAPTER 2. LITERATURE REVIEW

2.1 VQA and Domain Specialization

Visual Question Answering (VQA) is a complex task that necessitates the integration of computer vision and natural language processing techniques. It challenges systems to not only understand the visual content of an image but also to comprehend natural language questions related to that image and provide accurate answers. The majority of VQA research focuses on developing general-purpose models capable of handling a wide variety of images and questions**.** However, domain-specific VQA, which tailors' models to specific types of images and questions, can significantly improve performance by leveraging specialized knowledge and. For example, in medical imaging, VQA systems can assist doctors in diagnosing diseases by answering questions about X-rays or MRI scans. In robotics, VQA can enable robots to understand human instructions and interact with their environment. The relative scarcity of research directly addressing VQA for fruit and vegetable imagery highlights the need for a specialized system and motivates the development of this project. By focusing on this domain, we can exploit its unique characteristics to build a more accurate and efficient VQA system.

2.2 Image Feature Extraction with ResNet18

Convolutional Neural Networks (CNNs) have become the workhorse for image feature extraction in various computer vision tasks, including VQA. CNNs are capable of automatically learning hierarchical representations of images, capturing both low-level features (e.g., edges, textures) and high-level semantic concepts (e.g., objects, scenes). Among the various CNN architectures, pre-trained CNNs, such as ResNet18**,** offer significant advantages due to transfer learning. Transfer learning allows us to leverage the knowledge gained from training on a large dataset (e.g., ImageNet) and apply it to a new, smaller dataset. ResNet18, in particular, is a deep residual network with 18 layers. Its skip connections, or residual connections, mitigate the vanishing gradient problem, enabling effective training of deeper

networks. This architecture allows the model to learn more complex and abstract features from the images, leading to improved performance on the VQA task. The use of a pre-trained ResNet18 also reduces the training time and the need for large amounts of labeled data.

2.3 Question Encoding with BertTokenizer and LSTM

Natural Language Processing (NLP) techniques play a crucial role in understanding and processing questions in VQA. To effectively encode the textual information, we use the BertTokenizer to convert the question into a sequence of numerical tokens. BertTokenizer is a subword tokenizer based on the WordPiece algorithm. It breaks down words into smaller subwords, allowing the model to handle out-of-vocabulary words and capture the morphological structure of words. These token IDs are then fed into an embedding layer that maps them to dense vectors. To capture the sequential nature of language, we employ a Long Short-Term Memory (LSTM) network**.** LSTMs are a type of recurrent neural network (RNN) that are well-suited for processing sequential data. They have memory cells and gates that allow them to selectively remember and forget information, enabling them to capture long-range dependencies in the input sequence. The LSTM processes the sequence of word embeddings and generates a fixed-length vector representation of the question, capturing its semantic meaning.

2.4 Data Generation with BLIP

Creating large, labeled datasets for VQA is a challenging and time-consuming process. To address this challenge, we employ a pre-trained vision-language model, BLIP (Bootstrapping Language-Image Pre-training)**.** BLIP is a transformer-based model that is pre-trained on a large dataset of image-text pairs. It can be used for various vision-language tasks, including image captioning and visual question answering. In our project, we use BLIP to generate image captions for the fruit and vegetable images. These captions serve as descriptions of the images and can be used to create question-answer pairs. By leveraging BLIP, we can significantly reduce the

manual effort required for data annotation and create a more diverse and representative dataset for training our VQA model.

2.5 Training Optimization: AMP and Gradient Accumulation

Training deep learning models, particularly large models like ResNet18 and LSTMs, can be computationally expensive and require significant GPU memory. To address these challenges, we employ two optimization techniques: mixed precision training (AMP) and gradient accumulation. Mixed precision training (AMP speeds up the training process and reduces memory consumption by using lower-precision floating-point numbers (e.g., FP16) for certain operations. This can lead to significant speedups without sacrificing accuracy. Gradient accumulation allows us to simulate larger batch sizes when limited by GPU memory. Instead of updating the model parameters after each mini-batch, we accumulate the gradients over multiple mini-batches before performing an update. This effectively increases the batch size and can lead to improved training stability and generalization performance.

While significant progress has been made in the field of VQA, several challenges remain. General-purpose VQA models often struggle with domain-specific tasks where specialized knowledge is required. Furthermore, the creation of large, labeled datasets for VQA is a time-consuming and expensive process. This project addresses these gaps by exploring a domain-specific VQA task (fruit and vegetable imagery) and using a pre-trained BLIP model for data generation. We also evaluate the impact of mixed precision training (AMP) and gradient accumulation on the training efficiency and performance of our VQA model. By combining these techniques, we aim to build a more accurate, efficient, and scalable VQA system for the fruit and vegetable domain.

CHAPTER 3. METHODOLOGY

3.1 Dataset and Preprocessing

1.1 Dataset description

Accurate classification of fruits and vegetables is essential for applications in dietary monitoring, automated checkout systems, and food safety. Deep learning models, particularly Convolutional Neural Networks (CNNs) combined with Natural Language Processing (NLP), have demonstrated significant improvements in image-based food recognition. The *Fruit and Vegetable Image Recognition* dataset provides a well-structured benchmark for training such models, particularly for Visual Question Answering (VQA) applications, where the system must interpret and respond to textual queries about food images.

The dataset consists of high-quality images of 36 distinct fruit and vegetable classes, divided into training and test sets. Each class represents a specific type of food, allowing the dataset to be used for both classification and object recognition tasks.

The *Fruit and Vegetable Image Recognition* dataset was chosen for its diversity, structured labeling, and compatibility with deep learning architectures.

The *Fruit and Vegetable Image Recognition* dataset serves as a high-quality benchmark for image classification and VQA applications in food recognition. Its structured organization, diverse categories, and real-world applicability make it an ideal choice for training deep learning models. Future research can extend this dataset by incorporating object detection techniques and integrating it with large-scale multi-modal datasets for improved performance in automated food recognition systems.

1.2 Question - Answer Pair Generation

To train the VQA model, a dataset of question-answer pairs was created. The questions were generated using a combination of two methods:

Template-Based Question Generation: A set of question templates was defined to cover common question types [list a few examples: "What fruit is this?", "Is this a fruit or vegetable?". These templates were then populated with specific fruit and vegetable names from the dataset.

BLIP-Based Question Generation: The BLIP model was used to generate captions for each image. These captions were then used as a basis for generating questions. This approach allowed for the generation of more diverse and contextually relevant questions. The answers were manually verified and corrected to ensure accuracy.

1.3 Data Preprocessing

Image Preprocessing: Images were resized to size 224x224 pixels and normalized using the standard ImageNet statistics. The following transformations were applied for data augmentation during training. The torchvision.transforms library was used to perform these transformations.

Text Preprocessing: Questions were tokenized using the BertTokenizer with a maximum sequence length. The tokenizer was used to convert the questions into numerical token IDs and create attention masks.

3.2 Model Architecture

1.1 Overview

The VQA model consists of three main components: an image feature extractor, a question encoder, and a multimodal fusion module. The image feature extractor is a pretrained ResNet18 CNN. The question encoder is an LSTM network that processes tokenized questions from BertTokenizer. The multimodal fusion module concatenates the image and question features and feeds them into a fully connected layer for answer prediction.

1.2 Image Feature Extraction

A pre-trained ResNet18 model was used for image feature extraction. The ResNet18 model was pre-trained on the ImageNet dataset and fine-tuned on the VQA dataset. The output of the ResNet18 model is a feature vector representing the image.

1.3 Question Encoding

Questions were encoded using an LSTM network. The input to the LSTM is a sequence of token IDs generated by the BertTokenizer. The LSTM has layers and a hidden size. The output of the LSTM is a vector representing the question.

1.4 Multimodal Fusion and Answer prediction

The image features and question embeddings are concatenated to create a combined representation. This combined representation is then fed into a fully connected layer with hidden units and a ReLU activation function. The output of the fully connected layer is fed into another fully connected layer with output units and a Softmax activation function to predict the answer probabilities.

3.3 Training Procedure and Optimization

The VQA model was trained using the Adam optimizer**.** The model was trained for 10 epochs. A StepLR learning rate scheduler was used with a step size of 5 and a gamma of 0.1.

Mixed Precision Training (AMP): Mixed precision training was enabled using the torch.cuda.amp module. The GradScaler was used to prevent underflow and overflow during training.

The training loop consisted of the following steps:

1. Forward pass through the model to obtain the predicted output.
2. Calculation of the cross-entropy loss between the predicted output and the ground truth labels.
3. Backward pass to compute the gradients.
4. Application of gradient scaling (if using AMP).

5. Accumulation of gradients over multiple mini-batches (if using gradient accumulation).
6. Updating the model parameters using the optimizer.
7. Updating the learning rate scheduler.
8. Evaluation Metrics

CHAPTER 4. EXPERIMENTS AND RESULTS

4.1 Results and Analysis

Optimizers are used to update weight parameters to minimize the loss function. Several types of optimizers, including SGD, Adagrad, Adadelta, and Adam, were tested with different learning rates. Among them, the Adam optimizer demonstrated the best performance, outperforming the others in terms of efficiency and accuracy.

The VQA model achieved an accuracy of **86 %** on the test set.

4.2 Discussion

The results demonstrate the effectiveness of the proposed VQA system for answering questions about fruit and vegetable imagery. The high accuracy achieved on the test set indicates that the model has learned to effectively integrate visual and textual information. The analysis of performance by question type reveals that the model performs better on some types of questions than others, suggesting that further improvements could be made by focusing on specific question types. The implementation of mixed precision training (AMP) significantly improved training efficiency without sacrificing accuracy. This allowed us to train the model faster and with less memory, making it more feasible to experiment with different architectures and hyperparameters. The use of gradient accumulation also proved to be beneficial, allowing us to effectively train with a larger batch size, which led to improved generalization performance. The qualitative analysis provides valuable insights into the model's strengths and weaknesses. The model is able to correctly answer questions about simple images, but struggles with more complex images or questions that require more nuanced reasoning. The error analysis suggests that future work could focus on improving the model's ability to distinguish between similar fruits and vegetables and to handle more complex questions.

CHAPTER 5. CONCLUSION

5.1 Summary and Limitations

This project successfully developed a Visual Question Answering (VQA) system tailored to the domain of fruit and vegetable imagery. The system combines a ResNet18 CNN for image feature extraction with an LSTM network for question encoding, and achieves an accuracy of **86**% on a held-out test set. A pre-trained BLIP model was effectively used to generate question-answer pairs, reducing the need for manual annotation and enabling the creation of a more diverse training dataset. Furthermore, the implementation of mixed precision training (AMP) and gradient accumulation significantly improved training efficiency, allowing for faster experimentation and reduced memory consumption.

However, the project also has several limitations:

Dataset Size: The dataset, while sufficient for demonstrating the feasibility of the approach, is relatively small compared to those used in general-purpose VQA research. A larger dataset could potentially lead to improved model performance and generalization.

Question Types: The question types considered in this project were limited to a predefined set. The model may not generalize well to more complex or open-ended questions.

Model Complexity: While ResNet18 and LSTM provide a solid foundation, more complex model architectures (e.g., Transformers) could potentially capture more intricate relationships between images and questions. However, these models typically require more computational resources and larger datasets.

BLIP-Generated Data Quality: While BLIP was useful for data generation, the generated questions and answers were not always perfect and required manual verification. The quality of the BLIP-generated data may have limited the overall performance of the VQA system.

Limited Experimentation: Due to time and resource constraints, not all possible hyperparameter combinations or model architectures could be explored.

5.2 Future Work

The limitations of this project suggest several promising directions for future research:

- **Expanding the Dataset:** Increasing the size and diversity of the dataset by incorporating more images and generating a wider range of question-answer pairs. This could involve using more sophisticated data augmentation techniques or exploring alternative data generation methods. Using data from multiple sources and combining them.
- **Exploring More Complex Model Architectures:** Investigating the use of more powerful models, such as Transformers, to improve the accuracy and generalization performance of the VQA system. This would require significant computational resources and careful tuning of hyperparameters.
- **Improving Data Generation with Better LLMs:** Implementing few-shot learning and RLHF in training LLMs for data generation. Also, exploring alternative data generation methods that can produce more accurate and diverse question-answer pairs. This could involve using different pre-trained models or developing new algorithms for generating questions and answers from images.
- **Incorporating External Knowledge:** Integrating external knowledge sources, such as knowledge graphs or ontologies, to provide the model with more contextual information and improve its ability to reason about the relationships between fruits and vegetables.
- **Addressing Open-Ended Questions:** Developing techniques to handle more complex or open-ended questions that require the model to generate more elaborate and nuanced answers. This could involve using techniques from

natural language generation or incorporating common-sense reasoning capabilities.

- **Expanding the Domain:** Extending the VQA system to other domains, such as medical imaging, robotics, or e-commerce. This would require adapting the model architecture, training procedure, and data generation techniques to the specific characteristics of the new domain.
- **Real-world testing:** Deploying the system in a real-world setting (e.g., a mobile app or a retail store) to evaluate its performance and identify areas for improvement.

5.3 Concluding Remarks

This project has demonstrated the feasibility of building a VQA system for the fruit and vegetable domain using deep learning techniques and a pre-trained vision-language model. The results show that the system can achieve reasonable accuracy on a limited set of question types. While several limitations remain, the project provides a solid foundation for future research in this area. By addressing these limitations and exploring the suggested directions for future work, it may be possible to create a VQA system that can reliably and accurately answer a wide range of questions about fruits and vegetables, with a wide range of real-world applications.

REFERENCES

[1] Zhangyang Wang, Shiyu Chang, Jiayu Zhou, Meng Wang, Thomas S. Huang, “Learning A Task-Specific Deep Architecture For Clustering,” 2015.

[2] Linting Xue, Collin F. Lynch, “Incorporating Task-specific Features into Deep Models to,” 2020.

[3] Waseem RawatZenghui, WangZenghui Wang, “Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review,” 2017.

[4] Alex KrizhevskyIlya, SutskeverGeoffrey, E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” 2012.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, “Deep Residual Learning for Image Recognition”.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” 2019, p. 4171–4186.

[7] Jurgen Schmidhuber, Sepp Hochreiter, “LONG SHORTTERM MEMORY,” 1997.

[8] Huiling, LiYan-Gao, HuoYan-Gao, HuoXi He, “Li et al-2022-Nature,” 2022.

[9] Wonyeol Lee, Rahul Sharma, Alex Aiken, “Training with Mixed-Precision Floating-Point Assignments,” 2023.

[10] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, Devi Parikh, “VQA: Visual Question Answering,” 27 Oct 2015. [Trực tuyến]. Available: https://arxiv.org/abs/1505.00468.

[11] “Image Classification on ImageNet,” [Trực tuyến]. Available: https://paperswithcode.com/sota/image-classification-on-imagenet.

[12] Diederik Kingma, Jimmy Ba, “Adam: A Method for Stochastic Optimization,” 2014.