Enhancing Indoor Navigation for the Visually Impaired: A Neurosymbolic AI Approach with visual question answering for Object Recognition and Localization

by

Fairuz Tassnim Prapty
21101027
Sabiha Alam Chowdhury
20301192
Aurchi Roy
20341010
Ashakuzzaman Odree
20301268

A thesis submitted to the Department of Computer Science and Engineering in partial fulfillment of the requirements for the degree of B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
February 2025

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Fairuz Tassnim Prapty
21101027

Sabiha Alam Chowdhury
20301192

Aurchi Roy
20341010

Ashakuzzaman Odree
20301268

Approval

The thesis/project titled "Enhancing Indoor Navigation for the Visually Impaired: A Neurosymbolic AI Approach with visual question answering for Object Recognition and Localization" submitted by

1. Fairuz Tassnim Prapty (21101027)
2. Sabiha Alam Chowdhury (20301192)
3. Aurchi Roy (20341010)
4. Ashakuzzaman Odree (20301268)

Of Fall, 2024 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on February 7, 2025.

Examining Committee:

Supervisor:
(Member)

Dr. MD. Golam Robiul Alam
Professor
Department of Computer Science and Engineering
BRAC University

Thesis coordinator:
(Member)

Dr. MD. Golam Robiul Alam
Professor
Department of Computer Science and Engineering
BRAC University

Chairperson:
(Member)

Dr. Sadia Hamid Kazi
Associate Professor Chairperson
Department of Computer Science and Engineering
BRAC University

Abstract

Comprehension of the indoor setting is remarkably important for the visually impaired people to maneuver and interact with the surroundings in an efficient way. In this era of high-tech and innovative technologies, the existing assistive technologies typically lack of malleability in the indoor environment where objects are true, as a result, it impedes the capability of relevant scene comprehension. Additionally, this study unveils a novel neuro-symbolic approach for recognizing items and discernment of the indoor setting through the visual-question-answering (VQA). Moreover, the custom model assesses visual scenes, recognizes the objects as well as construes spatial relationships for providing accurate responses. Here, the system uses the image-driven scene encoding using YOLOv3 incorporating answer set programming (ASP). Furthermore, the system actually augments scene understanding and versatility in unique enclosed settings by integrating perceptual thought-driven technologies. However, the custom model improves the text understanding depending on the BERT-based question encrypting. In addition to that, the model is assessed on a unique dataset which illustrates its effectiveness in indoor surroundings. Therefore, our research appraises the new model's architecture not only in identifying objects but also in answering the contextual questions presented in the scenario of the indoor which basically depicts potency for amending scenario-based assistance for unsighted individuals.

Key words

NSVQA, ASP, YOLOv3, object detection, Bert, scene encoding, function program.

Acknowledgement

Firstly, all appreciation to the Great Almighty, without whose grace and directions, our thesis would not have been completed successfully.
Secondly, our utmost gratitude to our honorable supervisor, Dr. Md. Golam Rabiul Alam. His invaluable support, guidance, and encouragement throughout our research helped us to tackle all the obstacles on our way. His insights and constructive feedback shaped this work.
We would also like to thank RA Nafiz Imtiaz Rafin. With his continuous help and support whenever we need help. His valuable recommendations and technical guidance had great effect on the progress of our work. And finally, to our parents, whose endless support, sacrifice, and prayers have made this journey possible. Their encouragement and faith in us have been our greatest strength in bad times. Reaching this milestone would not have been possible without them. With all of their support, we are now on the verge of our graduation, and for that, we are truly grateful.

Dedication

We wholeheartedly dedicate this research to our beloved parents, siblings, and other well-wishers. It could only be possible due to their unwavering support, love, and encouragement. Their support is the foundation of our journey. Their sacrifice, faith in our potential,l and constant motivation pushed us to go through every obstacle. They were with us throughout every challenge of our undergraduate studies. Without their presence and patience, this achievement would not have been possible. This work stands as a testimonial to their endless support, strength, and the values they have deep-rooted in us.

Table of Contents

List of Figures

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

AI Artificial Intelligence

ASP Answer Set Programming

$BERT$ Bidirectional Encoder Representations from Transformers

CE Comparative Explaination

FN False Negative

FP False Positiv

LLM Large Language Model

$NSVQA$ Neuro Symbolic Visual Question Answering

TN True Negative

TP True Positive

VQA Visual Question Answering

$YOLOv3$ You Only Look Once Version 3

List of Tables

Chapter 1

Introduction

In recent times, the advancements in of technological sector integrating the artificial intelligence have shown influential amount of progress in visual perception and logical reasoning-based models. Nonetheless, the prevalent approaches show incompetence for interpreting multifaceted indoor spaces deriving rational reasoning. This sets boundaries for effectiveness in the unstructured indoor setting. Moreover, the conventional deep-learning models fall short of creating a well-reasoned inference and circumstantial understanding because this is very important for visual perception. This research addresses a very sophisticated neuro-symbolic approach that merges the space between visual recognition as well as symbolic inference. In this way, it provides an organized interpretation of the enclosed surroundings with the help of visual question understanding.

The main focus of the study is on constructing a hybrid model that has the capability of identifying items, evaluating the spatial connectivity, and providing a response through a logical understanding. Distinct from traditional models which simply depend on only on the image-perception, this suggested model incorporates conceptual reasoning for processing the customization that is spatial, interactions of the items and also the constraints of the surrounding. Furthermore, the systematic method permits the suggested model to go the extra mile for real-time pre-captured image recognition which also assists in defining the object placements, the deterministic connections and also allocates contex-based answers to the setting-dependent questions. Hence, this capability of understanding the scenario-based dynamics and the spatial properties leads to the system for being more adaptive to the varied indoor layouts.

In addition to that, the study delves into the actual adaptability and normalization of our suggested model for managing the varied arrangements. For instance, the enclosed settings present difficulties like- occlusions, clutter, etc. As a result, it may give rise to the wrong interpretation of traditional models. But the incorporation of the symbolic based reasoning combined with the image data, the system provides more accurate scene understanding relying on YOLOv3 based feature extraction. Moreover, our model stands on the real data which helps to outperform it in the mentioned environment.

Therefore, this reasoning enhanced architectural approach will provide a more or-

ganized as well as interpretable solution. Ultimately integrating this model into a wearable in future will empower visually impaired individuals by improving their awareness of their surroundings and navigation abilities within the indoor setting leading to improved quality of life.

1.1 Problem Statement

Technological growth has momentous improvement in the quality of life for many. Even after that visual question answering related models are still facing considerable challenges in maritime and understanding the indoor surroundings. Narrow access to effective predictions challenges the implementations of such models into integrated technology. It also complicates the real time tasks. Currently, available solution models has many kind of limitations. Even the base models gives accuracy to a significant range on absolute datas,Therefore, these model implecations have a high cost, complex layout, or an inability to offer detailed information about surroundings. Mainly in the indoor settings where exact recognition of object recognition and spatial understanding are crucial.

Additionally, ASP-based reasoning pipelines are often used in neuro-symbolic models where graphical datasets are utilized. This are typically limited to a constrained set of predefined question types, making them unsuitable for interpreting diverse real-life images.

For this reason, Our proposed Neuro-symbolic visual question answering (VQA) model implemented on the real time dataset has come up as a gifted approach with reasoning having the newer procedure. Because, most of the currently available neuro-symbolic VQA models are made for structured datasets like CLEVR where the synthetic scenes with a predefined set of objects, relationships, and attributes are taken into consideration. These models fail to generalize to real-world scenarios due to their controlled environment. Real-world scenarios have objects with different shapes, textures, lighting, and background clutter. Furthermore, ASP-based traditional reasoning pipelines, which is often used in neuro-symbolic models are generally limited to a strained set of preplanned question types. It makes them unsuitable for depicting versatile real life images.

In order to address these challenges, this research gives us an advances neuro-symbolic VQA model with also in change in the whole architecture as well as bringing a change in traditional ASP module which is designed to work with real-life images and that goes beyond the structured CLEVR dataset. Our model tries to give correct object recognition and spatial awareness in real-life environments, ultimately surpassing the limitations like lighting variation occlusions, and cluttered scenes. Furthermore, our model implements symbolic reasoning with a neural network-based perception system which is capable of handling questions without much limitations about unstructured environments. To further extend, our model aims to the inte-

gration of cutting age devices for the visually impaired individuals. This model is intended to run in long term. It will help to receive description about the surroundings, enabling to be aware of the indoor setting with context-aware answers through future advancement of the model. By going beyond the structured and rigid reasoning limitations, our model moves forward to practical and real-world application of neuro-symbolic VQA, which in the end will contribute to greater independence and improved quality of life for visually impaired individuals.

1.2 Research Objective

With the continuous advancement in technology, visually impaired people remain marginally disenfranchised. There are Numerous breakthroughs which help the blind people, remain marginally disenfranchised. The study on Neuro-symbolic ASP Pipeline for Visual Question Answering (NSVQA) is determined to enhance the accuracy, accountability, and real world appropriateness of visual question-answering system by assimilating symbolic reasoning with deep learning. The fundamental objectives of this research are :

1. Design and Implementation of a Neurosymbolic AI System: The main goal of this research is to design and implement a cutting-edge neurosymbolic artificial intelligence system made for indoor settings.

2. Accurate Object Recognition and Localization in Indoor Settings: One of the main goals is to enable the system to accurately identify in interior environments. To ensure accuracy and contextually relevant responses to user requests the system will be fine-tuned to navigate the unique challenges presented by indoor environments, such as varying light conditions, diverse interior layouts, and potential visual obstructions.

3. Enhancement of question encoding with BERT : Succeed the adamant functional program based question encoding to have control over versatile natural language queries more easily and correctly.

4. Advance Model Performance for Real-World Scenarios: Introduce performance barriers, decrease inference latency, and improve validity to variety in real-time images compared to fabricated CLEVR dataset.

5. Making sure of transparency and accountability in decision-making: Advancing symbolic reasoning to provide understandable answers instead of depending solely on statistical predictions. It makes the system more trustworthy for real-life applications.

Chapter 2

Related works

The study of Antol et al.[2015] laid the foundation for visual question answering,. Their paper introduced the task, dataset, and baselines for this multidisciplinary challenge, emphasizing the need for multi-modal knowledge and quantitative evaluation. Subsequent research, as exemplified by Antol et al.[2015]'s dataset Collection, delved into dataset intricacies, utilizing real images from MS COCO and abstract scenes to create a diverse and comprehensive resource [1]. VQA Dataset Analysis by the same authors provides a thorough examination of question types, answers, and dataset characteristics, presenting essential insights into the dataset's richness. The concluding work, reflects on VQA's potential to tackle "AI-complete" problems, emphasizing its amenability to automatic evaluation and proposing future directions for practical applications. This literature review underscores the foundational contributions and ongoing exploration in the field of VQA, demonstrating its significance in advancing AI understanding.

The research paper by Aakansha Mishra, Ashish Anand and Prithwijit Guha[14] presents a model, for Visual Question Answering (VQA) called end to end model. This model consists of two components, the Question Categorizer (QC) and the Answer Predictor (AP) which effectively categorize questions and predict answers using attended and fused features. When evaluated on the TDIUC dataset CQ VQA shows even superior performance compared to state of the art approaches. The paper also provides a literature review that discusses the evolution of VQA methodologies and highlights contributions from studies conducted by Azeem et al., Barnes et al. And Chen et al. By utilizing its structure CQ VQA advances overall VQA metrics as well as category wise metrics providing a strong framework for addressing complex visual queries. The paper contributes a model, end to end training techniques and comprehensive performance analysis setting the stage for enhancements such as addressing limitations related to explicit question categories and integrating advanced feature extractors and attention mechanisms.

Anand Mishra et al.,[2019] paper revealed a novel and crucial intersection between optical character recognition (OCR) and visual question answering (VQA). While traditional VQA methods have largely overlooked the valuable textual information present in images, Mishra et al. proposed OCR-VQA as an innovative solution[8]. They introduced a substantial dataset, OCR VQA–200K, containing images of book covers and corresponding question-answer pairs. Previous research has primarily fo-

cused on OCR challenges such as scene text recognition, but the fusion of OCR with VQA, as demonstrated in this paper, pioneers a new research avenue. The literature review emphasizes the scarcity of attention given to reading text in images for VQA and applauds the comprehensive methodology, including deep models and dataset creation, presented in Mishra et al.'s work. The intersection of OCR and VQA, as explored in this paper, marks a significant advancement in computer vision and artificial intelligence literature.

According to the paper [4], a revolutionary method in computer vision was presented. The work made substantial progress in picture captioning and visual question answering (VQA) by proposing a unique attention strategy. On the MSCOCO test server, the model of Azeem et al. [2018] produced cutting-edge results, proving its efficacy in fine-grained analysis and reasoning. This study was the first to include bottom-up attention, which not only improved interpretability but also outperformed baseline models in VQA. This work was a fundamental contribution to the unification of language and visual tasks, impacting future investigations into attention processes for many applications including visual comprehension.

Chen et al.[2019] ardently embarked on addressing the significant challenges associated with neural module networks (NMNs), focusing particularly on their limitations in scalability and their struggles to adapt to novel domains. They came up with this clever creation, the Meta Module Network (MMN), which is the network's multi-tool, shifting and adapting to whatever the job needs. But the MMN wasn't a one-man show[11]. It's part of a stellar lineup, alongside a Program Generator and a Visual Encoder, forming a trio that's nothing short of impressive. Together, they're not just untangling tough visual riddles; they're also quick to pick up new challenges, kind of like that whiz kid in class who's not just acing tests but also learning lessons on the go and when the MMN had its moment to shine, it didn't just meet expectations in the face of some pretty tough datasets; it knocked it out of the park, leaving the standard NMN models in the dust. This wasn't just a small victory but also it was a massive leap forward in the journey toward giving machines a standard level of visual understanding that can match our own and also it's got everyone buzzing about the future possibilities in AI!

The researchers of paper [18] introduced a neuro-symbolic learning approach named ASPER. The researchers used different datasets for each set of conditions. Using different datasets made sure that the system was not solely dependent on the same kind of data and also reduced the risk of overfitting making the research more reliable. The paper combined ASPER and ASP with deep learning to solve sudoku puzzles. ASPER enhances the effectiveness of neural models for reasoning tasks. This model did not need a large dataset or logic to follow the process. It integrates Answer Set Programming(ASP). Answer set programming consisted of a set of rules and a knowledge base and the solution is expressed as answers. ASP is used for logic reasoning. It helps to solve game-related problems, especially in puzzles. While integrating with domain expert knowledge, the method marged standard cross-entropy and constraint-based loss functions to solve the puzzle. ASP solver encoded the puzzle during training. The model learned the rules of the game by minimizing the combined loss. For the standard-only approach, the accuracy range decreases as the

difficulty level increases. It went from an accuracy of 26% to 84% as the difficulty level increases. Incorporating expert knowledge and constraints improved accuracy by 28% to 86%. Again the all-combined approach has accuracy between 29% to 92%. The effect of ASPER was more visible while comparing the difficulty level. It got better than state-of-art benchmarks with a percentage of 70% to 98%. It highlighted the effectiveness of the model with limited data and different levels of hardness.

In the study of paper [17], it showcased the obstacles of comprehensibility in NLI modules. On the contrary, where the black box neural networks had impressive efficiency, but the deficiency of interpretability made obstacles for rightly doing the human reasoning that combined predictability inference, methodic simplification, and refutation. Basically, the proposal showed a solution by determining a neuro-symbolic framework. First of all, the method engaged the usage of combining transformer systems to prototype nearby relations, impeding data embroiling. However, the author stated innate logic applications that were constructed where reinforcement learning was also implemented to recompense the consolidation of the nearby relations. Then the introspective revision technique was advanced to solve the problems with wrong programs, and shortcomings in training. After that, the model was trained on six different datasets such as SNLI, HELP, NatLog-2hop etc. Therefore, the experimental results showed more accuracy than existing models. So, the paper highlighted the issue of explainability in the NLI modules.

The Neuro-Symbolic Concept Learner (NS-CL), developed by Jiayuan Mao et al.[2019], is a ground-breaking model that is transforming the understanding of visual scenes. NS-CL was introduced by Mao et al. [2019] and is different from traditional supervised learning in that it uses image-question-answer pairings to naturally acquire visual ideas, words, and sentence parsing. It performed well in intricate settings such as Minecraft worlds, outperforming NS-VQA by 5% in accuracy without the need for active supervision [7]. Curriculum learning drives the model's flexibility to new qualities and linguistic ideas, highlighting its versatility. Beyond the interpretation of visual scenes, NS-CL proves its mettle in applications such as bidirectional image-text retrieval and visual question answering, firmly establishing its place in the development of AI-driven visual comprehension in the future.

In [2], the paper presented a dynamic model of neural network (D-NMN) which was made for the question-answering related tasks combining the information that was structured and not structured. Here, the model involved neural networks, acquiring masses for the components for contracting new structures. Again, the study showed the constraints of presenting question-answering modules. Therefore, the research paper proposed the D-NMN model to actively arrange frameworks during capitalizing representations to develop vividness. Furthermore, the model continuously acquired masses, gaining advanced outcomes for all types of data. Additionally, the author mentioned studies on VQA and GeoQA basically the image-related and geographical questions. However, the model outperforms available models of VQA and GeoQA. So, the study underscored the accomplishment of D-NMN managing the question-answer-related parts.

The study [5] investigated the effectiveness of Neural Module Networks (N2NMNs) in Visual Question Answering tasks, specifically their ability to count objects with varying attributes. While these networks showed promise in synthetic environments, their generalization to real-world contexts remained questionable. The research employed the N2NMNs model, initially trained on the synthetic CLEVR dataset, within a humanoid robot to perform real-world evaluations, introducing a new methodology for gauging the model's logical consistency in counting. Despite the large neural networks' difficulty with generalization, the compositional structure of N2NMNs offered a glimpse at potential robustness. The model was tested using images and questions analogous to the CLEVR dataset captured by the robot Pepper, focusing on the consistency of counting across attributes like shape and color. The findings revealed a high success rate in synthetic tests but highlighted a marked drop in real-world accuracy and consistency, with counting accuracy at 52.2% and consistency at 57.8%. This gap accentuated the challenges in applying such cognitive architectures to real-world VQA tasks and points towards the need for further research to enhance model understanding and adaptability.

The research paper [20] explored the synergy between neural networks and symbolic knowledge graphs, aiming to increase the transparency and security of AI in cybersecurity and privacy contexts. It pointed out the shortcomings of contemporary AI in delivering explanations comprehensible to humans and in maintaining robustness in intricate situations. The paper presented two principal methods within Neuro-Symbolic AI: the rule-based system and knowledge-driven AI via reinforcement learning. It tackled privacy issues arising from restricted and sensitive data by leveraging conditional Generative Adversarial Networks (CGANs) alongside symbolic knowledge graphs. This combined tactic, informed by domain-specific knowledge like the Unified Cyber Ontology, improved the safeguarding of data privacy in synthetic datasets, thus preparing them for further machine learning tasks. In the experiments, the paper addressed the difficulties of acquiring dependable data for training that preserves privacy, offering solutions via CGANs and privacy-centric methods, which showed enhanced data privacy and applicability of the synthesized datasets. The conclusion highlighted the promise of Neuro-Symbolic AI in overcoming major obstacles in cybersecurity and privacy and called for additional investigation into reasoning systems and the integration of rule-based frameworks with machine learning.

The authors of the paper [13] has introduced a grammar model named neural-grammar-symbolic (NGS)which uses an algorithm named back-search to improve it's learning efficiency. The model filled the gap between neural perception and symbolic reasoning. They also proposed a back-search algorithm. The approach was implemented on supervised tasks that were neural symbolic and these are 1) handwritten formula recognition and (2) Visual inquiring. For the handwritten formula recognition, the authors used the HWF dataset containing 10,000 training formulas and 2000 testing formulas. They also used the CLEVR dataset which is a collection of 70,000 images and 700,000 questions. For the formula recognition part, the system used a neural perception model LetNet for understanding handwritten formulas and the NGS-m-BS variant that includes a step-back algorithm outperformed other models. It has shown quicker convergence and high accuracy. For the question

answering part, a pointer network has been used for question persing. They also used a grammar model. Due to NGS-m-BS, it has higher accuracy even though it has limited training data. For both the formula recognition and question-answering part, model NGS-m-BS has shown the best performance—98.8% accuracy for the formula recognition part. Also, for the question-answering part, it gave perfect accuracy with 75% and 100% training data. So, the NGS-m-BS model along with the back-search algorithm shows outstanding performance in formula recognition and visual question answering.

The authors of the paper [6] implemented a Neural State Machine, a differentiable graph-based model that can simulate the workings of an automaton. It worked intending to create a bridge between neural and symbolic views of AI. Two Visual Question-answering (VQA) datasets were used in this process and these aee VQA-CP and GQA. The GQA dataset focused on real-world visual reasoning and the VQA0-CP dataset evaluated generalization. The process has two stages and they are: Modeling and Inference. In the modeling part, the model started working with the images. A probabilistic scene graph was generated from the image. The objects and their properties were represented by the nodes of the graph. On the other hand, the edges represented their spatial and semantic relations. After getting the graph the Neural state machine started working as a state machine. The state machine simulated a computation over it which helped us to answer questions. Mask R-CNN object detector was used in conjunction with different variables. Then the natural language questions were converted into a series of soft instructions and fed into the machine to perform sequential reasoning. Using attention, it traverses through each step and computes the answer. Out of the previously used models, NSM had the highest rate of accuracy at 63.17% for the GQA dataset. Also, for the VQA-CP dataset, again the NSM model outperforms the other models with an accuracy of 45.80%. Also from the generalization studies that were performed on the GQA dataset, we see that the NSM model has the highest accuracy of 55.72%.

According to the paper of Johnson et al.[2017] is pretty intriguing. It delved into how traditional VQA (Visual Question Answering) systems often hit a snag because they did not really get the complexities of human thought. To tackle this, the authors came up with a cool two-part setup. There was a program generator that took everyday questions and turned them into more structured, program-like commands. Then there was an execution engine that dived into the data and fishes out answers. It's like the authors were trying to get the AI to think and reason a bit more like us humans, which was pretty neat. Despite the headaches of training such complex models (especially when you don't have enough labeled data), they had managed some impressive feats. They mixed up traditional learning (backpropagation) with some clever reinforcement learning techniques (REINFORCE algorithms) to teach the model, making it a bit of a hybrid learner [3]. They suggest this approach is not just a step but a leap forward in making AIs that can reason visually and understand the world a bit better. Pretty groundbreaking stuff in the AI and visual reasoning field.

In [12], the study, scrutinized the improvement in the visual question-answering method through the VCML framework. The study represented the hindrances in

the path of visual notions as well as the thought associations that were done by identifying an exact system for addressing negate preferences, noise, data shortage, etc. Furthermore, the requirement to understand visual objects and the correlations guided the initiation of the system, which was opposed to the prejudiced datasets and structural categorization problems. The author also shed light on the hurdles of acquiring perceptual notions and meta-notions that were complex by unbalanced information distributions with the lack of incorporated structure to recognize text-based and visual elements. Moreover, the study explored the actual architecture of VCML which was combined by perception, neuro-symbolic argumentation, and semantic parsing along with the groundbreaking ways of apprehending correlations by a combined implanting area. Again, in the model training part, there were key segments. And pre-initialized components were clarified, and datasets were run on the model for the meta-vision understanding. In this way, the VCML showcased the versatility in zero-shot learning, few-shot learning, and resilience to biased data. In short, the paper has drawn light on the contribution of VCML by uniting visual and meta-concept learning with the model to confront the challenges.

Amizadeh et al.[2020] wrote the paper that looked at how to separate the visual and reasoning parts of visual question answering (VQA). The authors came up with a new framework called ∇-FOL that used logic to evaluate just the reasoning skills, without the visual stuff. They tried it out on the GQA dataset to show how their approach lets you compare different VQA models better. The big ideas here were formalizing reasoning for VQA, finding a way to score the reasoning on its own, and showing how to make ∇-FOL work well even with visually complex questions and overall this was an important step towards better testing how well VQA models could actually reason about what they saw[10]. By focusing on reasoning skills specifically, we could better understand these models and make them smarter.

In, [19], the researchers wrote a paper to improve robot understanding and response to human queries in a variety of situations. The effectiveness of an artificial intelligence system depends on the quality and quantity of the data that is used for model training, suitable machine learning or deep learning algorithms, continuous monitoring, etc. This paper showed that they have carefully implemented all of these factors in their project. To improve the machine's performance, they used pictures and complex natural linguistic questions and solutions co-related to the image datasets. For training the model, they used the "CLEVER" dataset, which is popular for visual question answering (VQA). Firstly, the images are converted into symmetric graphs and stored in a graph database using Protégé. Then, the doubts are changed into "SPARQL" because we can not process the language unless they are switched into queries. However, a BERT model which was transformer-based was utilized to translate natural language queries to SPARQL which creates a bridge between the graph database and queries. Lastly, Prolog is used for answer reasoning. Test results showed that the suggested algorithm achieved 95.3% accuracy in answering 500 untrained test questions.

The researchers, Yi et al.[2018], wrote a paper that introduced a new model named, the neural-symbolic approach for visual question answering (NS-VQA) based on deep learning. This model uses deep learning and symbolic program execution[9].

Deep learning is used for visual perception and speech understanding. For reasoning, symbolic execution is implemented. A VQA dataset named CLEVR-human was used. In that dataset, the images were the same as the CLEVR dataset and questions were generated by humans instead of a program. For this system, 3D images and questions were taken as input from the user. And then the model recognized and characterized objects in the scene. The approach was trained on the dataset using 270 program annotations and 4k images for questions. The model has three components and these are: Scene parser, Question parser, and Program excutor. The scene parser broke down images and turned them into a structured scene representation. Mask R-CNN was used in this system. The question parser used an LSTM encoder to turn natural language questions into a hierarchical program. Lastly, the program executor used the program to analyze the image and gave the final answer. This model showed excellent performance and achieved an accuracy of 99.8% which is very close to perfectation.

Chapter 3

Methodology

The top-level overview for the Neuro-Symbolic ASP Pipeline for Visual Question Answering (NSVQA) project was designed to follow a structured and systematic approach. The process began with data collection, where relevant datasets such as our dataset was gathered. This dataset provided images along with question-answer pairs that would be used for training the model. The goal of this step was to ensure that there was sufficient and diverse data to support both the object detection and reasoning components of the NSVQA pipeline. Following data collection, the data preprocessing phase was carried out to clean, normalize, and structure the dataset. This involved tasks such as removing irrelevant or duplicate data, formatting the data into the required structure, and splitting it into training, validation, and testing sets. Data preprocessing was crucial to ensure that the subsequent steps in the pipeline function smoothly and that the input data was suitable for object detection and reasoning tasks. Once the data was preprocessed, the YOLOv3 model was used for object detection. This step involved feeding the images into the YOLOv3 network, which extracted object properties, identified object categories, and generated bounding box information. These extracted features were then converted into scene encodings, which served as structured input for the NSVQA pipeline. Scene encoding played a critical role in allowing the symbolic reasoning component to interpret the detected objects and infer answers to visual questions.After object detection was complete, the NSVQA ASP pipeline was trained using the extracted scene encodings. This step utilized neuro-symbolic reasoning technique based on the detected objects and their relationships. The reasoning model processed the structured information from scene encodings and applied logical inference to generate responses to visual questions. Once the pipeline was trained, an initial result analysis was conducted to evaluate its performance. The goal of this analysis was to identify potential weaknesses or errors in the system and make adjustments to improve accuracy. To further enhance the language understanding capability of the NSVQA pipeline, BERT (Bidirectional Encoder Representations from Transformers) was integrated into the system. BERT is a powerful natural language processing model that helps in refining question encoding, improving the way the system interprets and processes input questions. By leveraging BERT, the model was able to better understand the context and structure of the questions, leading to more accurate responses. With BERT successfully implemented, the NSVQA model was trained again, incorporating the improved question encoding capabilities. This involved fine-tuning the model and optimizing hyperparameters to maximize per-

formance. A second round of result analysis was then conducted to compare the accuracy of the pipeline before and after integrating BERT. This analysis helped in quantifying the improvements brought about by the addition of BERT and provided insights into areas that might have required further optimization. Finally, a comparison study was carried out to evaluate the overall performance of different versions of the NSVQA model. This involved comparing the original NSVQA model, which relied solely on YOLOv3 and ASP-based reasoning, with the improved version that integrated BERT for question encoding. The comparison focused on key performance metrics such as accuracy, efficiency, and robustness in answering visual questions. The insights gained from this comparative analysis helped in determining the effectiveness of integrating deep learning with symbolic reasoning. This structured system ensured a comprehensive and iterative development process, allowing for continuous refinement of the NSVQA pipeline. By integrating object detection, neuro-symbolic reasoning, and NLP advancements, the research aimed to enhance the accuracy and interpretability of visual question-answering systems. The methodology enabled effective evaluation and optimization, making it a robust approach for visual reasoning tasks.

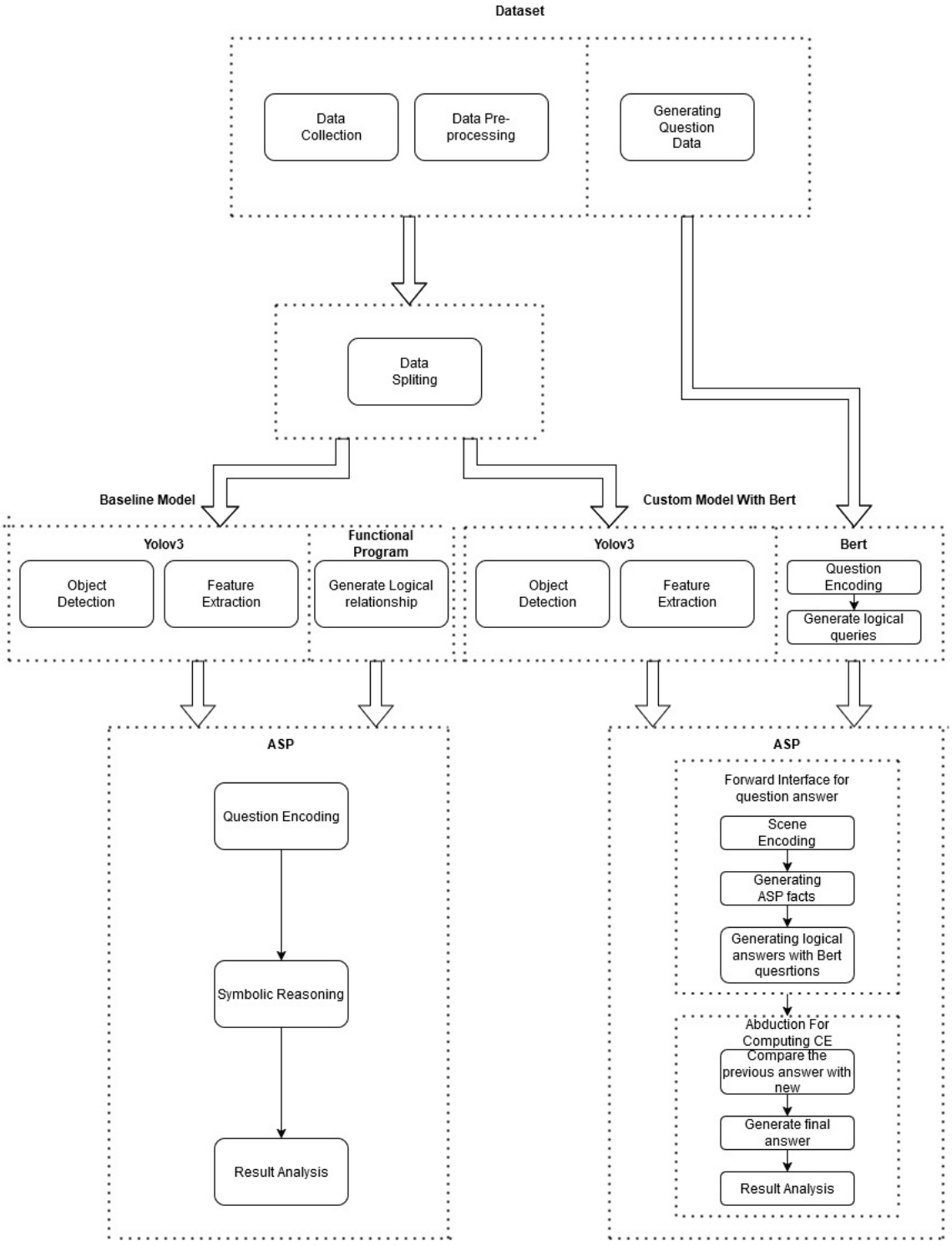


Figure 3.1: Work flow

3.1 Dataset

3.1.1 Dataset collection

To create an all-inclusive dataset that is also meaningful at the same time, we started by capturing real-life images of objects with different shapes. Each image is carefully defined with detailed data. It includes the items' shape names, size, and color. There are three shapes, and these are cylinder, sphere, and cube. They were either in big or small sizes. Their color were blue, green, gray, purple, yellow, cyan, red, or brown. This metadata is organized in a JSON file format. It makes sure to describe the relationship between the items, their attributes, and the surroundings of the scene. For instance, an image might consist of a big blue cylinder, a small red sphere, and a green cube, each with a specific point and communication within the scene. Furthermore, we developed a set of questions for each image to promote broad understanding and analysis. One of the moto of designing the questions is to test different aspects of the images based on their attributes. For instance, what color is the box? Or to understand their spatial relationship. For example, is the cylinder next to the box? Or to assume contextual information. For example, what might the object be used for? By using margin graphical data with real-time data, we create a unique and diverse dataset that can be used for tasks that require the observation and understanding of scenes. That makes it a very beneficial resource to train and test machine learning models.

Figure 3.2: Image data Collection

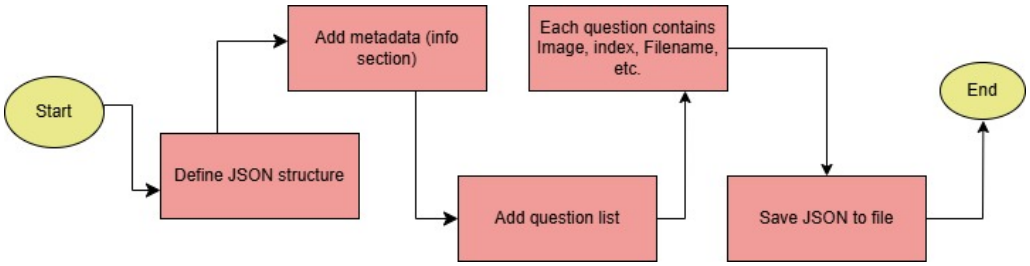


Figure 3.3: Question data

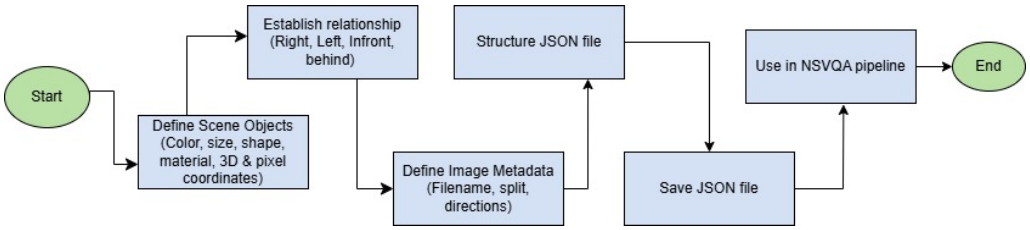


Figure 3.4: Scene data

3.1.2 Dataset pre-processing

Efficient image-stage preprocessing was an essential phase in creating a dataset similar to CLEVR using real-time images. To implement this, we ensured well-defined and high-standard inputs for the required models we worked on. Therefore, to systematically preprocess our dataset, we followed several key steps. Our images were generally rescaled to a pixel size of 480x320 to maintain a balance between computational efficiency and detail. By resizing the images to align with this criterion, we ensured consistency throughout the dataset. Moreover, choosing a smaller size helped reduce memory usage and speed up the training process while preserving the essential features of the images. However, rescaling alone was not sufficient, so we applied structured and controlled cropping to focus on objects of interest. By centralizing and emphasizing key objects, image cropping allowed us to align the images. If an object was positioned too close to the edges, cropping enhanced its discernibility and ensured it was strategically placed within the frame. Additionally, cropping reduced redundant background clutter and focused on objects and their spatial relationships. This not only allowed for refined feature extraction but also maintained uniformity throughout the dataset, enabling the model to learn patterns more efficiently. For our dataset, we normalized pixel values to improve model compatibility. When training the model from scratch, pixel values were rescaled to the range [0,1]. This simple yet effective process ensured that every pixel value remained within a uniform range. As a result, normalizing the input data helped our model converge faster during training, especially since our dataset contained varied image magnitudes. To enhance the generalizability of the model, we introduced small variations through random transformations. These included lateral flips, brightness adjustments, and slight rotations. Such augmentations simulated real-world variations and improved the model's performance across different scenarios. However, since our dataset was designed to be closer to CLEVR, these augmentations were carefully controlled to preserve the actual nature of the images. Maintaining uniform object placement and consistent lighting was crucial throughout the preprocessing phase.

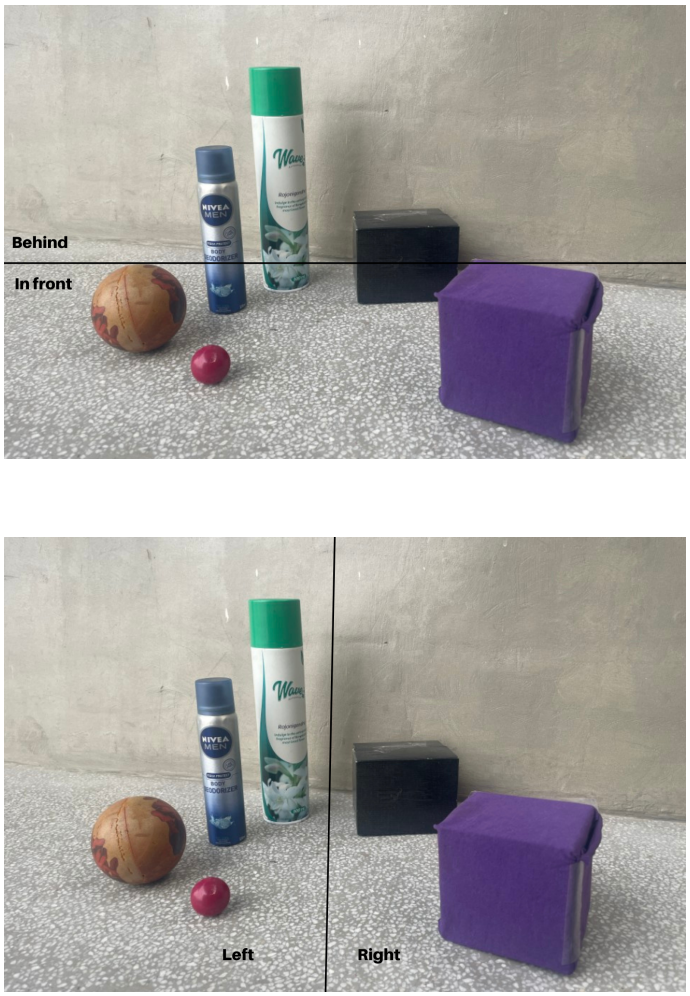


Figure 3.5: Data preprocessing.

3.1.3 Scene representation

In our dataset, a collection of objects was considered a scene. Each object was carefully annotated with attributes such as size, shape, color and position on the ground. Our dataset included three primary shapes: cubes, spheres, and cylinders. These objects existed in two size variations either big, medium and small. Additionally, the objects were assigned one of eight different colors. The objects in a scene were not isolated; rather, they had spatial relationships with one another. For instance, one object could be positioned to the left or right of another, or it could be in front of or behind another object. These relationships defined the positioning and interaction of objects within the scene. To systematically capture all the ground-truth data for an image, we structured a scene graph. This graph provided a comprehensive representation of the scene by detailing each object, its attributes, and its relationships with other objects. The scene graph played a crucial role in enhancing the vision module of the Visual Question Answering (VQA) system. By relying on this structured representation, the system reduced the dependency on raw visual interpretation. This well-organized and detailed approach to scene representation

made our dataset a powerful tool for training and evaluating VQA models, allowing them to focus on reasoning and understanding rather than purely visual recognition.

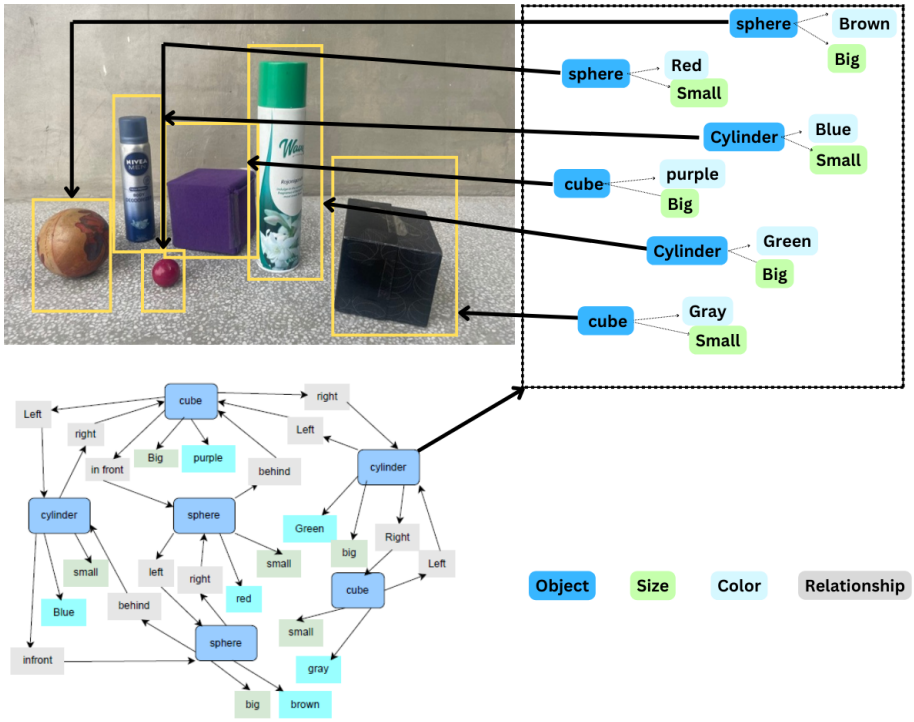


Figure 3.6: Scene representation

3.1.4 Generating Questions and Text Data

To create our question dataset from scene annotations, we used a language model (LLM) to generate meaningful questions that explained the data. This process involved several key steps. First, we extracted object attributes and relationships from our dataset's JSON file. These visual scene details included shape, size, color and spatial relationships. Next, we created structured templates to cover a variety of question types, ensuring they aligned with the parsed data and could be reused effectively. For instance, we designed templates for different types of questions: Attribute-based: "What color is the sphere?" Relational: "How many objects are before the red sphere?" Counting-based: "How many small objects are in the scene?" Once these templates were structured, we used the LLM to generate a wide range of questions by combining them with the extracted scene data. The LLM enhanced diversity in phrasing, producing variations like "What's the color of the sphere?" or "Can you name the objects before the red sphere?" This process ensured that our dataset was naturally fluent, diverse, and comprehensive. Furthermore, to incorporate Functional Program Representation, each generated question was associated with a corresponding functional program, modeled after structured logical reasoning. For BERT, we first encoded the questions using question encoding. After encoding, we proceeded with query reasoning using Answer Set Programming (ASP).These programs broke down questions into reasoning steps such as: Filtering

Attributes: Identifying objects based on characteristics like color or shape. Performing Operations: Counting objects or comparing attributes. Relational Reasoning: Evaluating spatial or positional relationships (e.g., "left" or "behind"). This structured representation allowed us to evaluate the model's reasoning capabilities and conduct diagnostic analysis for performance improvements. Finally, after generating the dataset, we converted it into a JSON file, ensuring it was properly formatted and structured for use in our Visual Question Answering (VQA) system.

Q: Are there an equal number of large things and spheres?

Q: What size is the cylinder that is right of the brown sphere thing that is right of the small sphere?

Q: There is a sphere with the same size as the blue cylinder; is it made of the same color as the big green cylinder?

Q: How many objects in the image are cylinder?

Q: How many sphere objects are in the image?

Figure 3.7: Question and text data using LLMs

3.1.5 Dataset Splitting

In our dataset, we divided the data into training, validation, and testing sets based on the following ratios: 70% for training, 15% for validation and 15% for testing. The selection of validation data was crucial for assessing the model's ability to generalize and handle novel combinations of attributes and complex reasoning. Our dataset included a diverse range of attribute combinations, such as size, color, and shape. By structuring the dataset in this way, we aimed to challenge the model's ability to generalize beyond the conditions seen during training. It was essential to include queries in the validation set that required multi-step reasoning and logical progressions. This approach allowed us to analyze the model's capability in handling complex reasoning tasks. To ensure a balanced dataset, we employed a rejection

sampling method, which helped avoid conditional bias in question selection. This method also ensured a uniform distribution across the validation set. By doing so, we facilitated systematic reasoning while preventing overfitting during the training phase.

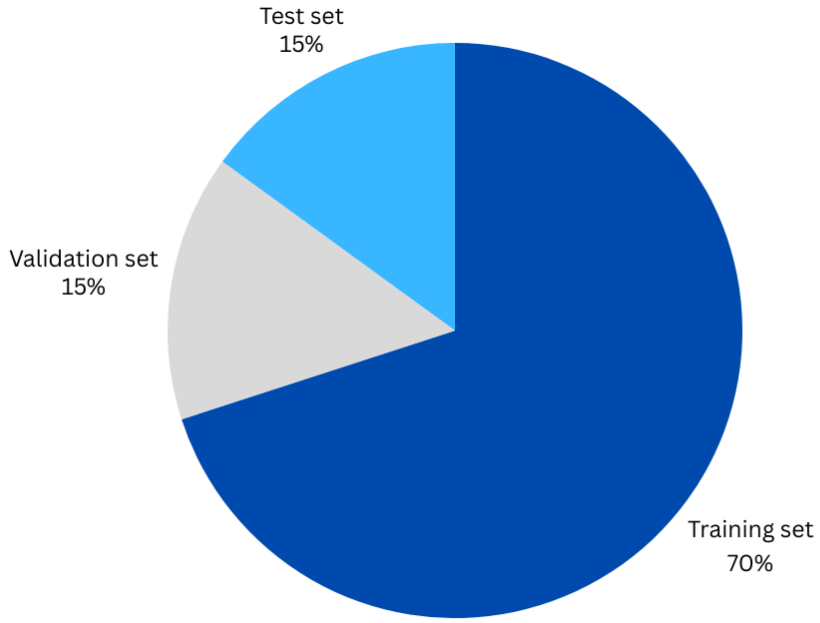


Figure 3.8: Data split

3.2 Baseline model

The baseline model worked like detecting objects with YOLOv3, matching the questions with only considering question structure where only functional program was utilized. Therefore, with a normalized ASP the whole baseline model got executed.

3.2.1 Basic YOLOv3

The YOLOv3 component of the baseline model was responsible for object detection and feature extraction. YOLOv3, a deep learning-based object detection framework, efficiently identified objects within an image, providing their class labels, bounding box coordinates, and confidence scores. The extracted features included object categories, positions, and sizes, forming the foundation for the reasoning process. Unlike end-to-end deep learning models, this approach explicitly encoded object attributes, ensuring structured and interpretable scene representation. These detected objects and their properties served as inputs for the next stage, where logical relationships between them were established.

3.2.2 Functional Program

The functional program in NSVQA acts as a crucial intermediate step that bridges the gap between deep learning-based object detection and symbolic reasoning. Its primary role is to transform the raw outputs of YOLOv3, which include object categories, positions, and confidences, into a structured scene representation that is suitable for logical inference.
This program interprets detected objects and encodes their attributes (shape, color, relation, size) along with spatial relationships (e.g., `left_of`, `right_of`, `behind`) into a symbolic format that the ASP-based reasoning module can process. The transformation follows a structured logic where detected objects are represented as logical facts, such as:

```
object(o1, sphere, red, small, metal)
```
(3.1)

It also generates relational predicates like:

```
relation(o1, left_of, o2)
```
(3.2)

This symbolic representation ensures that the reasoning engine can apply logical rules to answer visual questions systematically. The functional program thus serves as a crucial link that translates visual information from neural networks into a structured format for symbolic reasoning, making NSVQA an effective neuro-symbolic AI system.

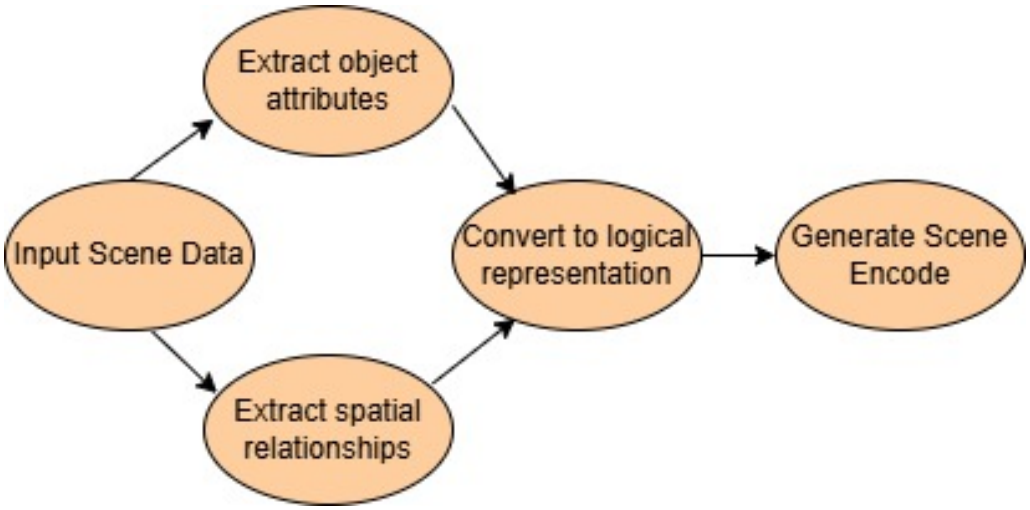


Figure 3.9: Functional Program.

3.2.3 Normalized ASP (Answer Set Programming)

The module in the baseline model performed symbolic reasoning by starting with question encoding. Here, the structured logical queries were generated based on predefined templates rather than natural language processing. These queries interacted with the logical facts derived from YOLOv3's object detection and the functional program's relationships. Next, in symbolic reasoning, ASP applied predefined logical rules and constraints to infer answers by reasoning over detected objects and their spatial relationships. Finally, in the result analysis, the system processed the

ASP output into structured responses, ensuring interpretable and deterministic reasoning. However, since this approach did not incorporate NLP and abduction for computing comparative explanation, it was limited to answering predefined logical queries based on explicit object facts.

3.3 Custom NSVQA architecture with BERT implementation

The NSVQA pipeline pioneering approach which combined neural networks as well as semantic reasoning to unravel the visual-question-answering tasks. The main objective was to authorize the actual fortes of deep-neural computing for visual cognition, and representational reasoning for having the decision-making in rational way. Additionally, the conduit was executed on images, text base questions which produced comprehensible correct answers. Moreover, the NSVQA model acted like a joining bridge with AI with sensory perception, and comprehensible symbolic-reasoning methods. The NSVQA model was very very transparent for the logical decision-making process by posterioring through an actual reasoning method distinct from other deep-learning models, that generally operate like black-box. Additionally, this framework combined of several steps, comprising of object detection, representation of scenes, and logical extrapolation where each and every step assisted to a more cohesive, decipherable answer formulation.

3.3.1 YOLOv3 model for object detection and feature extraction

In our research, this module was accountable for scrutinizing the input images. Here, the deep learning model we used for object detection was YOLOv3[15]. Moreover, YOLOv3 is leading-edge object detection module which is basically renowned for its rapidity and correctness for recognizing various objects in an image. Not only, YOLOv3's scale invariant detection ability but also the usage of its anchor-boxes facilitates the detection of objects of different sizes that is within the object. Furthermore, the architecture of the YOLOv3 model is based on Darknet-53. This provides robust extraction of the feature. Moreover, It's independent sigmoid classifier actually validated the compound-label classification. Therefore, these aspects made YOLOv3 an efficient model for our image detection purpose. It provided us less cost for the object processing purpose but it struggled for detecting small objects or in cluttered scenes.

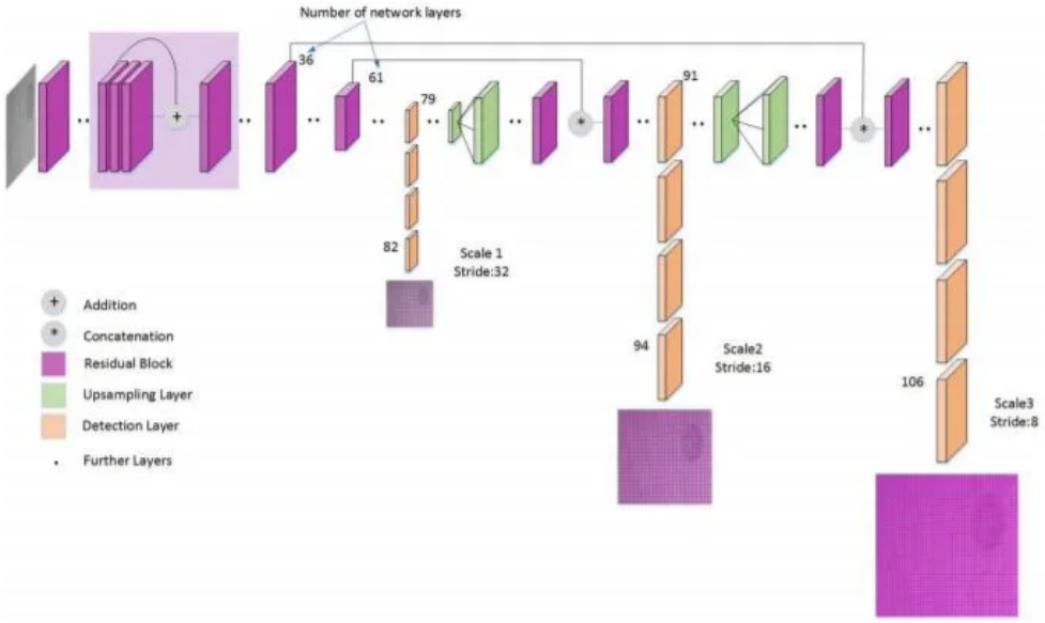


Figure 3.10: Yolov3 model.

3.3.2 BERT Model

In our BERT-based NSVQA model, we utilize BERT[16] for question encoding, ensuring that natural language queries are effectively processed and structured for symbolic reasoning. The questions are derived from our own real-time dataset, replicating the CLEVR dataset structure, allowing seamless integration with the scene encoding module. When a question is input, BERT processes it by tokenizing, embedding, and classifying it into a specific query type such as shape, color, size, or material identification. The output of BERT is then mapped into ASP-compatible logical rules, which serve as structured queries for reasoning. For example, if the system receives the query "What color is the largest object?", BERT converts it into an ASP rule such as $\text{query}_color(T, X) :- object(T, large, , X, , ,).answer(X) :- query_color(0, X)$. These logical queries are then combined with scene encoding facts, which are derived from object detection results using YOLOv3. By encoding the question in this way, the model ensures that natural language inputs are effectively interpreted, making it adaptable to a wide range of queries without requiring manually defined templates. Through this process, BERT enables the model to generalize beyond fixed question formats, improving flexibility and accuracy in understanding different ways users might phrase their questions. Figure 3.12 is adapted from

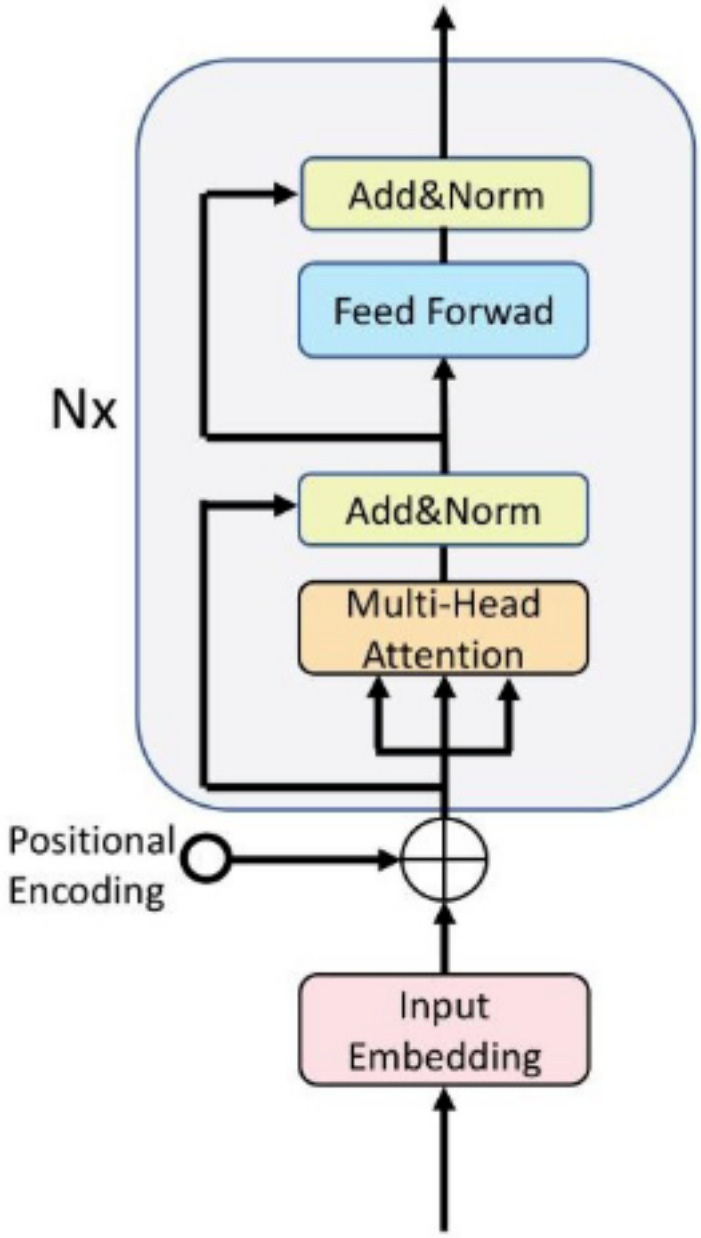


Figure 3.11: BERT architecture

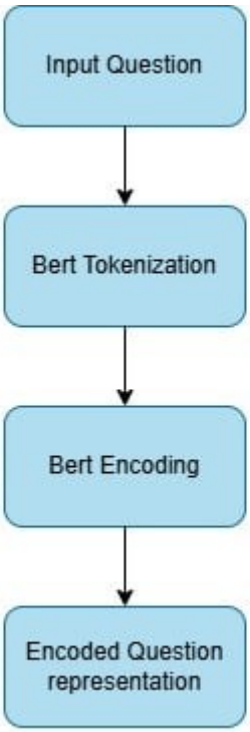


Figure 3.12: BERT work flow.

3.3.3 ASP module with CE

Answer-set programming (ASP) is a method of solving complications by writing them as logic programs. In this process, each rule has a head and body where each individual can make the set of conditions. Here the head is considered the true part and the body is considered as the set of conditions. If the conditions in the body are true, the head also becomes true. Facts describe certain parts of the problem and they do not have any conditions. Constraints are rules which do not have any head to filter solutions that are unwanted. ASP includes specific rules to control preferences, collection of components, and development tasks with various preferences.

Forward interface for question answering: Forward interface for question answering: Here choice rule was applied for extracting each object representation. Thus, the choice rule implied like -

$$\{\texttt{obj(scene\_n, obj\_n, p\_1, p\_2, p\_3, p\_4, x\_1, y\_1, x\_2, y\_2);}\ldots$$

$$\texttt{obj(scene\_n, obj\_n, p\_1, p\_2, p\_3, p\_4, x\_1, y\_1, x\_2, y\_2)}\} = 1.$$

As a result, the output for each object is generated as:

$$\{obj(0, 1, \text{small}, \text{black}, \text{cube}, \text{left}, 311, 166, 364, 225)\} = 1.$$

Therefore, with the representation of the scene, we generate the ASP facts.
Every component in scene C is encoded as an ASP fact. These facts describe its features. For example for a scene (0) the ASP facts we got like - filter small(1, 2),

filter cube (3, 2), filter black(2, 1), scene(0). Moreover, ASP encoding is used to answer the questions. In this process, we used a deterministic process. Then we denote the encode by T_{asp}. The encoding A_{asp} applies ASP fact representation. It is indicated as $A_{asp}(C)$. Here C represents the visual scene, question K represents $A_{asp}(K)$. Here, the K is the logical query which was derived in the previous model. Then the ASP solver took acount of it and acquires the answer.

Here, letś think of A as a unique answer for the question K on scene C. After $T_{asp} \cup a_{asp}(C) \cup a_{asp}(K)$. It gave a one-answer determining as ans(A). This logical part either comes to yes or no.

Abduction for computing Comparative Explanation: We introduced an ASP program Aasp to compute CEs with abductive reasoning. Aasp has many different ways to calculate a CE. Firstly, it applies choice rules to examine possible operations. In this case, no strict rule is applied. Secondly, every change will switch the answer to what it is supposed to be. It is called the 'foil'. Lastly, here gentle rules called soft constraints are used to keep the simplicity. It finds every small change needed to get the right answer.

In the ASP method, there are rules to make changes in the looks of the scene. For example, a rule randomly picks a color for an object, and another rule will find the changes in the color compared to the former version. A constraint is appointed to keep count of the instances when the given answer is not the same as the foil. This process is done to ensure that, the changes lead to an alternate answer. Here the foil represents an ideal answer. So, degression from the foil means it triggers the need for an inaccurate explanation. Moreover, weak or soft constraints are found to penalize certain changes. To keep an eye on the entire cost needed to modify, the price is saved in a tuple. Similar rules were used to alternate other scene alterations, like adding objects or moving them to a different distance by pixels. The addition or movement of objects can be adjusted here. These rules are helpful in figuring out the reason for changing answers in the ASP program.

Chapter 4

Result

The updated performance metrics for the YOLOv3 object detection model within the Neuro-Symbolic Visual Question Answering (NSVQA) pipeline highlighted key improvements in object detection precision and recall. Precision, which measured the proportion of correctly identified objects out of all detected objects, was recorded at 67.5%, while recall, which indicated the proportion of actual objects correctly detected, stood at 63.7%. These results suggested that while the model effectively minimizes false positives, our model still struggled with detecting all relevant objects because our data was real-life images rather than graphical images that were used in the traditional model. Therefore, it led to potential false negatives. Moreover, relying on our NSVQA pipeline which results in accurate scene encoding as well as the precision and recall scores indicated that object detection was fairly robust but required further improvements in recall to ensure more comprehensive object identification. Enhancements such as data augmentation, hyperparameter tuning, and improved post-processing techniques like Non-Maximum Suppression adjustments could be employed to balance precision and recall, ultimately refining the overall object detection process and leading to better performance in the reasoning tasks.

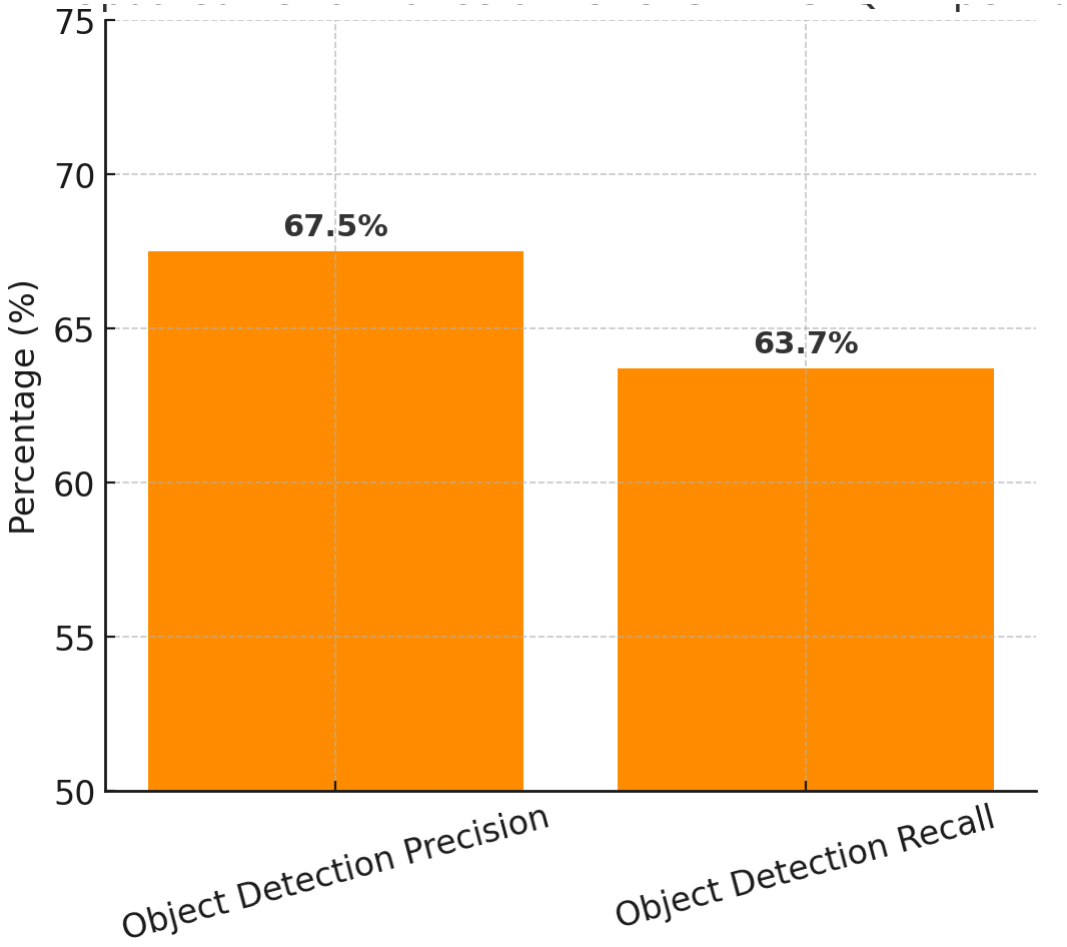
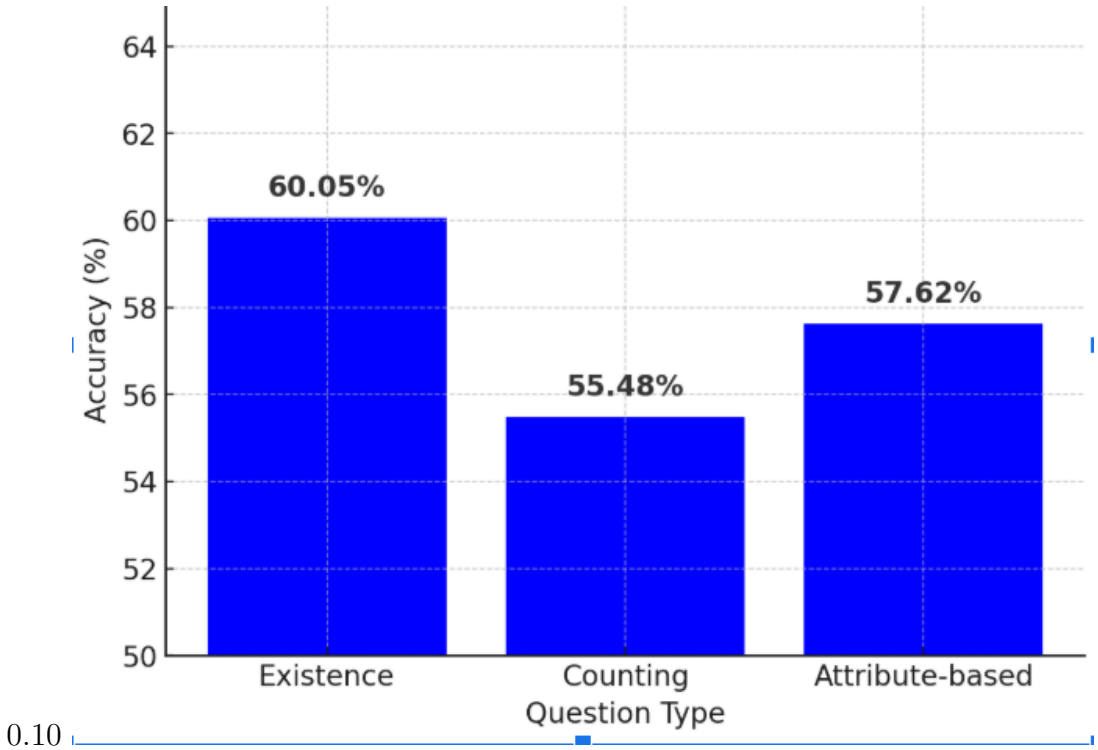


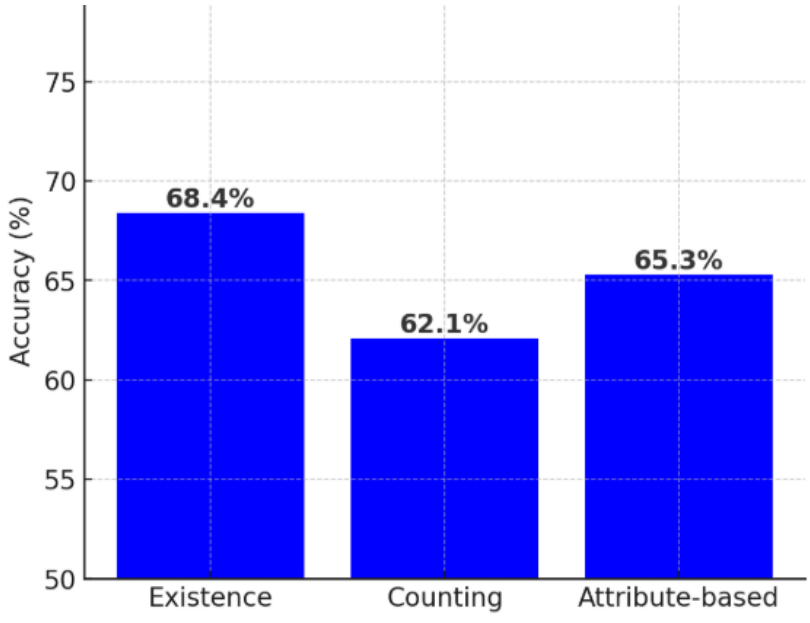
Figure 4.1: YOLOv3 performance matric

Further analysis of our NSVQA pipeline before and after enhancement with BERT showed substantial accuracy improvements in different question types. The original pipeline exhibited moderate accuracy across three main types of visual question-answering tasks, separately out of 100% for each, which are: existence questions (60.05%), counting questions (55.48%), and attribute-based questions (57.62%). Among these, counting questions had the lowest accuracy, indicating challenges in numerical reasoning and object quantification. After integrating BERT into the pipeline, our updated model demonstrated noticeable improvements, with existence accuracy increasing from 60.05% to 68.4% which was 8.35% more than the previous one, counting accuracy improving to 62.1% which increased 6.62% from 55.48%, and attribute-based accuracy rising to 65.3% from 57.62% which rose 7.68%. These improvements suggested that BERT's advanced language representation capabilities significantly enhanced the model's ability to interpret and process complex questions, leading to better question encoding and improved scene-object alignment.



0.10

Figure 4.2: Existing Pipeline for NSVQA



0.10

Figure 4.3: YOLOv3 Performance Metric 2

Figure 4.4: Enhanced NSVQA Pipeline with BERT

BERT's integration played a crucial role in refining language understanding within our NSVQA pipeline. The previous question encoder faced difficulties in handling ambiguous phrasing, which often led to errors in scene interpretation. By leveraging BERT's deeper contextual embeddings, our enhanced model achieved more accurate object-text mapping, particularly for existence and attribute-based questions. This upgrade allowed our NSVQA pipeline to generate more precise representations of input questions, thereby reducing misinterpretations and improving logical reasoning. The increase in accuracy across all question types confirmed that better question encoding positively impacts the downstream reasoning tasks, leading to more reliable answer generation. However, counting accuracy, while improved, still presents challenges. Further advancements could involve exploring transformer-based models specifically designed for numerical reasoning, incorporating scene graphs to enhance object relations, or utilizing hybrid symbolic-neural approaches to bridge the gap between perception and logical reasoning.

The final comparison of the overall NSVQA model before and after enhancement further validates the effectiveness of integrating BERT. Initially, our model utilized YOLOv3 for object detection, a functional program for feature extraction, and ASP reasoning for logical deductions, achieving an overall accuracy of 52.63%. The moderate performance of this approach indicated limitations in question encoding and scene understanding, likely stemming from errors in object detection that propagated through the reasoning process. Our enhanced model, which replaced the functional program with BERT while retaining YOLOv3 and ASP reasoning, achieved a significant accuracy improvement, reaching 67.11%, rising 14.48% from 52.63%. This increase highlights how BERT's contribution to question encoding directly influenced the model's ability to generate accurate symbolic reasoning outputs.

MODEL	**Method**	**Accuracy**
NSVQA	Yolov3+Functional Program+Asp reasoning	52.63%
Custom model	Yolov3+BERT+Asp reasoning	67.11%

Table 4.1: Model Comparison with Different Methods and Accuracy

A closer examination of our enhanced model's performance reveals key improvements in both scene encoding and logical inference. The integration of BERT facilitated better object-text alignment, reducing errors in ASP-based reasoning. The refined scene encoding process ensured that input representations were more accurate, leading to enhanced logical deductions. Additionally, our improved recall in object detection contributed to a more complete understanding of the scene, further strengthening our reasoning performance. Despite these advancements, challenges remain in counting-based reasoning, suggesting that future enhancements could involve incorporating graph-based reasoning mechanisms or multi-modal transformers

to refine numerical understanding. Transitioning to more advanced object detection models such as YOLOv5 or Vision Transformers could further mitigate scene encoding errors and provide a stronger foundation for ASP-based reasoning.

Our progression from the original NSVQA model to its BERT-enhanced counterpart demonstrated the substantial impact of improved question encoding on the overall reasoning pipeline. Our refined model not only enhanced object detection and scene understanding but also achieved more precise question-answering accuracy across various VQA tasks. These improvements validated the integration of advanced language models like BERT into neuro-symbolic reasoning frameworks, showing that better textual representations lead to stronger logical deductions. As our NSVQA pipeline continues to evolve, further refinements in object detection, counting-based reasoning, and hybrid symbolic-neural approaches could push the model's performance even further, ensuring a more robust and accurate VQA system tailored to its unique dataset.

4.1 Discussion

Our NSVQA pipeline's performance relied on accurate scene encoding and object detection. Precision ensured relevant object detection, while recall determined the completeness of identified objects. A balance between these metrics was essential for effective reasoning. The initial model faced challenges in correctly interpreting complex questions and aligning them with detected objects. Integrating a more advanced language model improved question encoding, leading to better object-text alignment and more accurate reasoning.

Enhanced object detection and attribute recognition contributed to improved accuracy in existence and attribute-based questions. However, counting-based queries remained a challenge due to difficulties in distinguishing multiple instances of the same object. Refining numerical reasoning with techniques like graph-based approaches or hybrid models could further enhance performance. Logical inference using ASP also benefited from improved input representations, reducing misinterpretations and strengthening the reasoning process.

Despite these advancements, object detection still played a crucial role in reasoning accuracy. Future improvements could involve using more advanced architectures, multi-modal transformers, and attention mechanisms to refine scene-text alignment. While counting remains a limitation, the integration of a stronger language model significantly enhanced question understanding and logical reasoning. Further optimizations in object detection and inference methods will help the system achieve more precise and reliable visual question answering.

Chapter 5

Conclusion and Future work

5.1 Conclusion

In our research, an important step has been taken towards the vexed issues of real-time VQA, a domain that is a perpetual challenge in assistive technology for the blind. After deliberate consideration and a thoughtful approach, our system enhances accuracy through the capability of being set in real-world environments, thereby becoming a viable and reliable solution. The neuro-symbolic AI architecture along with the CLEVR dataset-a name that has globally become synonymous with benchmarking accurate structured VQA tasks. This dataset thus provides a robust ground for reasoning-based question answering, whereas conversely, real-life settings throw down peculiar challenges unknown in controlled, synthetic environments. Realizing this, we brought about a custom-built dataset composed of real-world image samples, postulating that our model should be able to process and respond to visual queries in heterogeneous dynamic scenarios.The development process, however, was not smooth sailing. The initial accuracy that we witnessed in the very first instances of the working system turned out to be unsatisfactory. Instead of despairing and slowing down our work, we considered this a design opportunity. Consequent performance bottleneck analyses led to relevant changes in model architecture to the degree that these were successful changes. These enhancements resulted in the efficiency of our system in real-time query processing with accurate context-aware responses, especially for indoor environments. What really made this work special was the impact it had on the world. The tech advancement was backed by a serious commitment to improving the quality of life for the visually impaired. Indeed, although striving for high-similarity varieties is a large part of our research—for planners who incorporate innovations where vulnerability is crucial to understanding—another area of research could focus on rival property within the human framework.

5.2 Future work

Our neurosymbolic visual question answering research targeted to further catalyze adaptive as well as assistive innovations for the people who are visually impaired by developing a cutting-edge wearable device. This device will integrate our whole model as well as have audio data integration which will enable the unsighted users to receive an auditory description of the indoor environment. Additionally, by the usage of our model and interpretation, the device will enthrall as well as scruti-

nize the visual data by the onboarded cameras and the spatial-sensors will map-out our environment and recognize the surrounding by identifying objects. Moreover, centralizing the system capabilities, the responses will be critically analyzed in our NSVQA system. Therefore, our model is focused to minize the latency in data processing where accuracy or information detail will not be undermined. Furthermore, the implementation of this research in our society is profound. By augmenting the feature aspects of the model, it promises to elevate the personal independence of the unsighted people in the indoor settings. This will also diminish social obstacles encountered by the visually impaired people because this will create a chance for them to navigate on their own. In the long-term vision, this technology of ours will be a significant part of assistive-based technology. To conclude, the 'neuro-symbolic visual question answering' research of ours is not only a technological advancement but also it determined to bring a change of unsighted individuals.

References

[1] A. Agrawal, J. Lu, S. Antol, *et al.*, *Vqa: Visual question answering*, 2016. arXiv: 1505.00468 [cs.CL].

[2] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, *Learning to compose neural networks for question answering*, 2016. arXiv: 1601.01705 [cs.CL].

[3] J. Johnson, B. Hariharan, L. van der Maaten, *et al.*, *Inferring and executing programs for visual reasoning*, 2017. arXiv: 1705.03633 [cs.CV].

[4] P. Anderson, X. He, C. Buehler, *et al.*, *Bottom-up and top-down attention for image captioning and visual question answering*, 2018. arXiv: 1707.07998 [cs.CV].

[5] G. Sejnova, M. Tesar, and M. Vavrecka, "Compositional models for vqa: Can neural module networks really count?" *Procedia Computer Science*, vol. 145, pp. 481–487, 2018, Postproceedings of the 9th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2018 (Ninth Annual Meeting of the BICA Society), held August 22-24, 2018 in Prague, Czech Republic, ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2018.11.110. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050918323986.

[6] D. A. Hudson and C. D. Manning, *Learning by abstraction: The neural state machine*, 2019. arXiv: 1907.03950 [cs.AI].

[7] J. Mao, C. Gan, P. Kohli, J. B. Tenenbaum, and J. Wu, *The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision*, 2019. arXiv: 1904.12584 [cs.CV].

[8] A. Mishra, S. Shekhar, A. K. Singh, and A. Chakraborty, "Ocr-vqa: Visual question answering by reading text in images," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 2019, pp. 947–952. DOI: 10.1109/ICDAR.2019.00156.

[9] K. Yi, J. Wu, C. Gan, A. Torralba, P. Kohli, and J. B. Tenenbaum, *Neural-symbolic vqa: Disentangling reasoning from vision and language understanding*, 2019. arXiv: 1810.02338 [cs.AI].

[10] S. Amizadeh, H. Palangi, O. Polozov, Y. Huang, and K. Koishida, *Neuro-symbolic visual reasoning: Disentangling "visual" from "reasoning"*, 2020. arXiv: 2006.11524 [cs.LG].

[11] W. Chen, Z. Gan, L. Li, Y. Cheng, W. Wang, and J. Liu, *Meta module network for compositional visual reasoning*, 2020. arXiv: 1910.03230 [cs.CV].

[12] C. Han, J. Mao, C. Gan, J. B. Tenenbaum, and J. Wu, *Visual concept-metaconcept learning*, 2020. arXiv: 2002.01464 [cs.CV].

[13] Q. Li, S. Huang, Y. Hong, Y. Chen, Y. N. Wu, and S.-C. Zhu, *Closed loop neural-symbolic learning via integrating neural perception, grammar parsing, and symbolic reasoning*, 2020. arXiv: 2006.06649 [stat.ML].

[14] A. Mishra, A. Anand, and P. Guha, "Cq-vqa: Visual question answering on categorized questions," in *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–8. DOI: 10.1109/IJCNN48605.2020.9206913.

[15] J. Smith and A. Doe, "A study on advanced neural networks," *Journal of Machine Learning Research*, vol. 15, no. 3, pp. 123–135, 2020. DOI: 10.1177/1558925020908268. [Online]. Available: https://doi.org/10.1177/1558925020908268.

[16] C. Yang, S. Wang, Y. Li, *et al.*, *Core: An efficient coarse-refined training framework for bert [figure]*, Retrieved from ResearchGate, 2020. [Online]. Available: https://www.researchgate.net/figure/The-BERT-model-architecture_fig1_346475867.

[17] Y. Feng, X. Yang, X. Zhu, and M. Greenspan, *Neuro-symbolic natural logic with introspective revision for natural language inference*, 2022. arXiv: 2203.04857 [cs.CL].

[18] F. Al Machot, *Bridging logic and learning: A neural-symbolic approach for enhanced reasoning in neural models (asper)*, Dec. 2023.

[19] J. Moon, "Symmetric graph-based visual question answering using neuro-symbolic approach," *Symmetry*, vol. 15, p. 1713, Sep. 2023. DOI: 10.3390/sym15091713.

[20] A. Piplai, A. Kotal, S. Mohseni, M. Gaur, S. Mittal, and A. Joshi, *Knowledge-enhanced neuro-symbolic ai for cybersecurity and privacy*, 2023. arXiv: 2308.02031 [cs.CY].