YOLO based Visual Question Answering System for Pathology Images

A thesis submitted in partial fulfilment of the requirement for the

Degree of Master of Technology

Of

Jadavpur University

By

SOUGATA MAJUMDER

Registration Number: 154189 of 2020-2021

Examination Roll Number: M6TCT23012

Under the Guidance of

Dr. NIBARAN DAS

Professor

Department of Computer Science and Engineering

Jadavpur University, Kolkata-700032

India

June 2023

DECLARATION OF ORIGINALITY AND COMPLIANCE OF ACADEMIC ETHICS

I, Sougata Majumder, declare that the thesis titled, "YOLO based Visual Question Answering (VQA) for pathology images" and the work presented in it is my own. I confirm that:

- This work was done wholly or mainly while in candidature for master's degree at this University.
- Where any part of this thesis has previously been submitted for a degree or for any other qualifications at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is used as a reference. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all the primary sources of help.

Signature: _____
Date: _____

CERTIFICATE OF RECOMMENDATION

This is to certify that the work on this thesis entitled "YOLO based Visual Question Answering for pathology images" has been satisfactorily completed by Sougata Majumder, Examination Roll Number: M6TCT23012, University Registration No.: 154189 of 2020-2021. It is a bonafide piece of work carried out under my supervision at Jadavpur University, Kolkata-700032, for partial fulfillment of the requirements for the degree of Master of Technology in Computer Science and Engineering from the Department of Computer Science and Engineering, Jadavpur University for the academic session 2020-2023.

_____

Prof. (Dr.) Nibaran Das
(Supervisor)
Department of Computer Science & Engineering, Jadavpur University
Date: _____

_____

Prof. (Dr.) Nandini Mukhopadhyay
(Head of the Department)
Department of Computer Science & Engineering, Jadavpur University
Date: _____

_____

Prof. Ardhendu Ghoshal
(Dean)
Faculty of Engineering and Technology, Jadavpur University
Date: _____

Certificate of Approval

This is to certify that the thesis entitled ”YOLO based Visual Question Answering (VQA) for pathology images” is a bonafide record of work carried out by Sougata Majumder in fulfillment of the requirements for the award of the degree of Master of Engineering in Computer Science and Engineering from the Department of Computer Science and Engineering, Jadavpur University during the academic session 2020-2023. It is understood that by this approval the undersigned does not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein but approve the thesis only for the purpose for which it has been submitted.

Examiners:

_____
(Signature of Examiner)
Date: _____

_____
(Signature of Supervisor)
Date: _____

Acknowledgement

I would like to express my deepest gratitude to all those who have contributed to the completion of this thesis.

First and foremost, I am immensely grateful to my supervisor, Prof. Dr. NIBARAN DAS, for his guidance, expertise, and unwavering support throughout this research journey. Their invaluable insights and encouragement have played a pivotal role in shaping the direction and quality of this work.

I would also like to extend my appreciation to all the faculty members of Department of Computer Science and Engineering at Jadavpur University for their knowledge, inspiration, and valuable feedback. Their dedication to academic excellence has been instrumental in broadening my understanding of the subject matter. I am thankful to the members who provided their consent for the use of the medical images in this study. Their generosity and cooperation have made it possible to explore the potential of visual question answering in the context of cytology and histopathology. Furthermore, I would like to acknowledge the support and assistance received from my colleagues and friends. Their constructive discussions, technical assistance, and moral support have been immensely valuable throughout the research process. I am also indebted to the staff members of the library, laboratory, and administrative departments at Jadavpur University for their efficient and prompt assistance whenever needed. Last but not least, I would like to express my deepest gratitude to my family for their unwavering love, understanding, and encouragement. Their constant support and belief in my abilities have been the driving force behind my pursuit of knowledge and academic achievements. This work would not have been possible without the collective contributions and support of all those mentioned above. Thank you for being an integral part of this journey and for making a significant impact on the completion of this thesis.

Contents

List of Figures

List of Tables

List of Algorithms

List of Abbreviations

AI Artificial Intelligence. 2

BERT Bidirectional Encoder Representations from Transformers. 3

CNNs Convolutional Neural Networks. 8

CV Computer Vision. 2

MMQ Multiple Meta-model Quantifying. 15

NLP Natural Language Processing. 2

RNNs Recurrent Neural Networks. 8

RoBERTa Robustly Optimized BERT Pretraining Approach. 3

ROI Region of Interest. 31

Chapter 1

Introduction

Visual Question Answering (VQA) [Lin et al., 2023] in the medical domain is an emerging research area that effectively combines both Computer Vision (CV) and Artificial Intelligence (AI). It aims to develop models and systems that can automatically answer questions based on medical images along with their respective class labels.

In the medical field, VQA holds great potential for various applications. By analyzing medical images such as cytology and histopathology, VQA systems can be able to assist healthcare professionals in diagnosing diseases, interpreting test results, and providing personalized treatment recommendations. This technology can enhance the efficiency and accuracy of medical image analysis, leading to improved patient care and outcomes.

On the other hand, VQA [Al-Sadi et al., 2021] in the medical domain involves several challenges. Medical images are often complex and diverse, requiring advanced image processing techniques to extract meaningful information. Additionally, understanding and answering questions related to medical images requires deep knowledge integration of both visual and textual domains.

To tackle these challenges, researchers employ techniques such as image segmentation, object detection, and localization to extract relevant features from medical images. These features are then combined with Natural Language Processing (NLP) algorithms for generating accurate answers to questions about those images.

Advancements in deep learning, neural networks, and large-scale medical image datasets have fueled the progress of VQA in the medical domain. Researchers are actively working on developing robust VQA models specifically tailored for medical image analysis, leveraging techniques such as ensembled segmentation, object detection, and localization. These models are further enhanced by training and evaluating them using the

approaches like Bidirectional Encoder Representations from Transformers (BERT)-base or Robustly Optimized BERT Pretraining Approach (RoBERTa)-base NLP model architectures [Rothman and Gulli, 2022], which incorporate both visual and textual understanding.

Furthermore, the application of VQA in the medical domain extends beyond diagnostic support. It can also facilitate medical education and research by providing intelligent question-answering capabilities. VQA systems can serve as valuable educational tools, thereby allowing medical students and practitioners to interactively learn and gain insights from medical images. These systems can help in identifying anatomical structures, recognizing pathological conditions, and understanding the underlying mechanisms of diseases.

In the realm of medical research, VQA can contribute to large-scale data analysis and knowledge discovery. This can be done by automatically extracting information from medical images and answering research-related questions, VQA systems can assist researchers in identifying patterns, correlations, and novel insights that may otherwise be overlooked. This can pave the way for advancements in medical imaging techniques, disease understanding, and treatment strategies.

Additionally, the integration of VQA with other technologies holds significance for telemedicine applications. By combining VQA with remote imaging systems, healthcare providers can remotely analyze medical images and receive real-time answers to their questions. This enables timely decision-making, especially in situations where immediate access to specialized expertise is limited.

Moreover, the use of VQA in the medical domain opens up possibilities for patient engagement and personalized healthcare. This has been possible by involving patients in the question-answering process, VQA systems can empower individuals to better understand their medical conditions, treatment options, and follow-up care. This enhanced patient-system interaction fosters informed decision-making, improves patient satisfaction, and promotes active participation in healthcare management.

As the field of VQA in the medical domain continues to evolve, ongoing research focuses on addressing specific challenges and improving the performance of the models. This includes exploring novel techniques for multimodal fusion [Atrey et al., 2010], handling rare or domain-specific medical conditions, and adapting VQA models for real-time applications. Additionally, efforts are being made to create benchmark datasets and evaluation metrics tailored to the medical domain, enabling fair comparisons

and fostering advancements in the field.

The ultimate goal of VQA in the medical domain is to create intelligent systems that can effectively analyze medical images and provide insightful answers to complex medical questions. This research has the potential to revolutionize healthcare by assisting healthcare professionals in making accurate diagnoses, improving patient outcomes, and enabling more efficient medical decision-making processes.

In summary, the intersection of VQA, computer vision, and artificial intelligence in the medical domain offers immense potential for enhancing medical image analysis, education, research, telemedicine, and patient care. The ongoing research and development in this area are poised to revolutionize the healthcare landscape, leading to improved diagnostics, personalized treatment strategies, and better patient outcomes.

1.1 Applications of Visual Question Answering from medical images

Visual Question Answering (VQA) from medical images [He et al., 2020] has various applications across the healthcare domain. Some key applications are listed below:

- Diagnostic Assistance: VQA can assist healthcare professionals in diagnosing diseases by answering specific questions related to medical images. For example, doctors can ask questions about the presence of abnormalities, specific anatomical structures, or characteristics of a disease, and the VQA system can provide answers based on its analysis of the images. This helps in improving the accuracy and efficiency of disease diagnosis.

- Treatment Planning and Monitoring: VQA can aid in treatment planning and monitoring by answering questions about treatment options, surgical techniques, or post-treatment evaluation. Healthcare providers can seek guidance from VQA systems to determine the best course of action, assess treatment progress, or identify potential complications based on the analysis of medical images.

- Medical Education and Training: VQA systems can be used as educational tools to enhance medical education and training. Medical students, residents, and practitioners can interact with VQA systems to ask questions about medical images, enabling interactive learning and a deeper understanding of complex cases. VQA can help in teaching anatomy, pathology, and radiology, allowing learners to practice image interpretation and receive real-time feedback.

- Clinical Decision Support: VQA can serve as a clinical decision support tool by providing evidence-based answers to questions related to medical images. Healthcare professionals can consult VQA systems to validate their diagnostic hypotheses, confirm treatment decisions, or seek additional information for complex cases. This assists in improving the accuracy and reliability of clinical decisions.

- Telemedicine and Remote Consultations: VQA can facilitate telemedicine and remote consultations by enabling healthcare providers to remotely analyze medical images and receive answers to their questions in real-time. This is particularly valuable in scenarios where access to spe-

cialized expertise is limited, enabling timely and accurate decision-making for patients in remote or underserved areas.

- Research and Data Analysis: VQA systems can assist researchers in analyzing large-scale medical image datasets and extracting valuable insights. Researchers can ask questions about patterns, correlations, or specific image features, and the VQA system can provide answers based on its analysis of the images. This helps in accelerating medical research, discovering new knowledge, and advance the understanding of diseases.

- Patient Engagement and Health Literacy: VQA can play a significant role in patient engagement and health literacy. Patients can ask questions about their medical images, such as understanding their test results, visualizing their condition, or clarifying treatment procedures. VQA systems can provide clear and accessible answers, empowering patients to actively participate in their healthcare journey, make informed decisions, and improve their overall health literacy.

- Clinical Research and Trials: VQA from medical images can assist in clinical research and trials by answering queries related to participant selection, response evaluation, or image-based endpoints. Researchers can utilize VQA systems to extract quantitative and qualitative information from medical images, helping in the identification of suitable candidates for clinical trials, tracking treatment response, and assessing the effectiveness of novel interventions.

- Quality Assurance and Error Detection: VQA systems can contribute to quality assurance and error detection in medical imaging workflows. By answering questions about image quality, artifacts, or technical issues, VQA can help identify potential errors or inconsistencies in the acquisition, processing, or interpretation of medical images. This ensures the delivery of accurate and reliable results, reducing the chances of misdiagnosis or incorrect treatment decisions.

- Automated Reporting and Documentation: VQA from medical images can streamline the reporting and documentation process for healthcare professionals. By answering questions about image findings or measurements, VQA systems can generate automated reports that summarize the key information from medical images. This saves time and effort for clinicians, allowing them to focus on patient care and reducing the burden of manual report generation.

- Image Retrieval and Case-Based Reasoning: VQA can facilitate image retrieval and case-based reasoning in the medical domain. Healthcare professionals can ask questions about similar cases or specific image characteristics, and the VQA system can retrieve relevant medical images that match the query. This aids in decision-making, providing reference cases, and assisting in knowledge transfer among medical practitioners.

- Public Health and Population Studies: VQA from medical images can be applied in public health and population studies. Researchers and public health professionals can ask questions related to population-level trends, disease prevalence, or geographical variations in medical images. VQA systems can provide insights and answers based on the analysis of large-scale image datasets, assisting in public health planning, resource allocation, and disease surveillance.

  By harnessing the capabilities of VQA from medical images, these applications contribute to improved patient outcomes, enhanced medical knowledge, and more efficient healthcare practices. Overall, the applications of VQA from medical images are diverse and impactful, ranging from diagnostic assistance to education, research, and remote healthcare delivery. By leveraging the power of computer vision, artificial intelligence, and medical imaging, VQA systems contribute to improved patient care, enhanced medical decision-making, and advancements in medical knowledge. It opens up new possibilities for transforming the way medical images are analyzed, interpreted, and utilized in various healthcare sectors.

1.2 Overview of Visual Question Answering in Machine Learning

Visual Question Answering(VQA) is a fascinating research area that resides at the intersection of machine learning, computer vision, and artificial intelligence(AI). It involves developing models and algorithms, capable of understanding and answering questions about visual content, such as images or videos, using natural language processing techniques. VQA has gained significant attention in recent years due to its potential in order to bridge the gap between textual and visual information, enabling more interactive and human-like interactions with machines.

The goal of VQA is to create intelligent systems that can comprehend the content of visual data and generate accurate, aligned, and meaningful responses to questions posed in natural language. By combining computer vision algorithms with language understanding models, VQA systems can interpret the visual content and extract relevant information to generate appropriate answers.

The underlying techniques employed in VQA encompass a range of machine learning and AI approaches. Convolutional Neural Networks (CNNs) [Sarvamangala and Kulkarni, 2022] are commonly used for extracting visual features from images or video frames, enabling the understanding of visual content. Recurrent Neural Networks (RNNs) [Gaafar et al., 2022] or Transformer models are often employed for processing the textual input, capturing the contextual information, and generating relevant responses. Additionally, attention mechanisms are frequently utilized to focus on specific regions or objects in the visual data while processing the question.

VQA models are typically trained on large-scale datasets that contain pairs of images or videos along with corresponding questions and answers. These datasets are annotated by humans, providing ground truth answers for training the models. The models hereby learn to generalize from the training data and are then evaluated on separate test datasets to measure their performance in terms of accuracy, comprehensibility, and reasoning abilities.

The challenges in VQA arise from the complexity of both visual understanding and natural language processing. Visual content can be highly diverse and nuanced, encompassing a wide range of objects, scenes, and visual relationships. Understanding and answering questions about such content requires the models to possess robust visual perception capabilities. Additionally, the language understanding aspect involves handling

ambiguities, implicit context, and commonsense reasoning.

Advancements in deep learning, neural network architectures, and large-scale annotated datasets have significantly contributed to the progress of VQA in recent years. Researchers have explored various techniques to improve the performance of VQA models, including multi-modal fusion methods that effectively combine visual and textual information, attention mechanisms that focus on relevant regions or objects, and the integration of external knowledge sources for better reasoning abilities.

In the field of Visual Question Answering (VQA), Natural Language Processing(NLP) plays a crucial role in enabling machines to understand and generate human-like responses to questions about visual content. NLP techniques contribute significantly to the language processing aspect of VQA, allowing models to comprehend the meaning, context, and nuances of textual input.

NLP contributes to VQA in several ways:

- Question Understanding: NLP techniques are employed to process and understand the questions posed to the VQA system. These techniques involve tokenization, parsing, and syntactic analysis to extract the underlying structure and meaning of the questions. NLP models, such as Recurrent Neural Networks (RNNs) or Transformer models, are used to capture the semantic relationships and contextual information within the questions.

- Answer Generation: NLP techniques are used to generate appropriate answers based on the visual content and the questions. This involves text generation methods that leverage language models, sequence-to-sequence models, or template-based approaches. NLP models enable the VQA system to generate responses that are grammatically correct, coherent, and semantically aligned with the questions.

- Language Understanding: NLP models contribute to the understanding of natural language in the context of VQA. These models can handle lexical ambiguity, syntactic variations, and semantic nuances in the questions. By incorporating pre-trained language models, such as Bidirectional Encoder Representations from Transformers(BERT), the VQA system can effectively capture the contextual information and improve its comprehension of the questions.

- Reasoning and Inference: NLP techniques enable the VQA system to perform complex reasoning and inference based on textual input.

This involves techniques like semantic parsing, logical reasoning, and commonsense knowledge integration. NLP models assist in interpreting the questions, inferring implicit information, and making logical deductions to arrive at accurate answers.

- Language-Visual Fusion: NLP techniques facilitate the fusion of language and visual information in VQA. By combining the textual and visual modalities, NLP models allow the VQA system to reason about the visual content based on the language input. Attention mechanisms, which are commonly used in NLP, can be applied to focus on relevant regions or objects in the visual data while processing the textual information.

- Multi-modal Understanding: NLP techniques contribute to multi-modal understanding in VQA by integrating textual and visual features. This involves techniques like multi-modal fusion, where visual features and textual features are combined to obtain a comprehensive representation of the input. NLP models help in capturing the correlations and interactions between the textual and visual modalities, enabling a more holistic understanding of the question and the visual content.

The advancements in NLP, such as pre-trained language models, attention mechanisms, and contextual embeddings, have greatly improved the performance of VQA systems. These techniques enhance language understanding capabilities, reasoning abilities, and the overall accuracy and quality of the answers generated by VQA models.

In summary, NLP plays a fundamental role in Visual Question Answering by enabling machines to understand and generate responses to questions about visual content. It contributes to question understanding, answer generation, language understanding, reasoning, language-visual fusion, and multi-modal understanding. The combination of NLP and computer vision techniques in VQA leads to more effective and intelligent systems that can comprehend and respond to human queries in a visually rich context.

1.3 Motivation

The motivation for conducting research on Visual Question Answering (VQA) using unsupervised medical image datasets in the field of medical research stems from several key factors:

- Knowledge Discovery: Medical image datasets, particularly cytology and histopathology images, contain a wealth of valuable information that can aid in diagnosing diseases, understanding their progression, and identifying relevant biomarkers. VQA techniques applied to these datasets can help extract and uncover latent knowledge, patterns, and insights that might not be apparent through manual analysis alone. By posing questions to the VQA model, researchers can gain a deeper understanding of the visual content and potentially discover new correlations, trends, or characteristics that could contribute to medical research.

- Augmenting Diagnostic Capabilities: VQA models applied to medical image datasets have the potential to enhance diagnostic capabilities. By allowing healthcare professionals to ask questions about the visual content, VQA systems can provide additional insights and context to aid in accurate diagnosis and treatment planning. The integration of VQA in medical research can assist in improving the accuracy and efficiency of medical image interpretation, potentially leading to better patient outcomes and more effective healthcare interventions.

- Exploration of Unsupervised Learning: Unsupervised learning techniques, including unsupervised VQA, have gained attention in medical research due to the scarcity of annotated medical image datasets. Unlike supervised approaches that rely on labeled data, unsupervised VQA models can leverage the inherent structure and patterns within the medical image datasets without the need for explicit annotations. This motivates researchers to explore and develop unsupervised VQA methods to harness the potential of large-scale, unlabeled medical image repositories, opening up new avenues for knowledge extraction and analysis.

- Automated Analysis and Decision Support: VQA models applied to unsupervised medical image datasets can provide automated analysis and decision support. By generating answers to questions related to image content, VQA systems can assist healthcare professionals in

making informed decisions. This can range from assisting in identifying specific cell types or tissue structures to providing insights into disease characteristics, prognosis, or treatment options. VQA in medical research can contribute to more efficient and accurate analysis, enabling healthcare providers to make evidence-based decisions and improve patient care.

- Bridging the Gap between Visual and Textual Information: The integration of VQA in medical research helps bridge the gap between visual information present in medical images and textual information expressed through natural language queries. By enabling the interpretation and generation of questions and answers, VQA models facilitate more interactive and intuitive interactions with medical image data. This integration of visual and textual modalities promotes a deeper understanding of the visual content and fosters collaborative efforts between medical experts and machine learning techniques.

- Advancing AI-Assisted Healthcare: The application of VQA in medical research aligns with the broader goal of advancing AI-assisted healthcare. By leveraging machine learning techniques and medical image analysis, VQA models contribute to the development of intelligent systems that can aid healthcare professionals in decision-making, improve diagnostics, and enhance patient care. The motivation lies in harnessing the potential of AI to augment human expertise, improve efficiency, and ultimately transform healthcare delivery.

Overall the motivation for exploring VQA techniques using unsupervised medical image datasets in medical research stems from the desire to uncover hidden knowledge, augment diagnostic capabilities, explore unsupervised learning approaches, provide automated analysis and decision support, bridge the gap between visual and textual information, and advance AI-assisted healthcare. These motivations drive researchers to leverage VQA methods to unlock the full potential of medical image data and contribute to advancements in medical research and patient care.

1.4 Contribution

The contributions of this research work can be summarized as follows:

- **Novel Dataset Creation**: The dataset is manually annotated, providing it with class labels or image features. This dataset serves as a valuable resource for training and evaluating the proposed system, enabling the development of accurate and contextually relevant question-answering capabilities.
- **Integration of YOLOv5 and Image Encoders**: The research work incorporates the YOLOv5 architecture for object detection, leveraging its precise bounding box predictions and class labels. Additionally, image encoders and CNN models are employed to extract visual features from the detected objects, enhancing the understanding of the image content.
- **Integration of Language Encoders and Fusion Algorithm**: The research work integrates language encoders and a fusion algorithm with the visual features extracted from the image. This integration enables a comprehensive understanding of both visual and textual modalities, facilitating the generation of meaningful questions and answers.
- **Question and Answer Generation**: The proposed system utilizes the combined visual and textual information to generate relevant questions based on the detected objects' class labels or image features. This process involves leveraging NLP networks to generate contextually relevant and meaningful questions.
- **Dataset Utilization**: The researcher employs a manually created JSON dataset to fetch the question and answer pairs for training and evaluation. This dataset, with its annotated class labels or image features, serves as a crucial resource in training the system and evaluating its performance.

Overall, the research work contributes to the field by creating a novel dataset with manual annotations, integrating YOLOv5 with image encoders along with language encoders and fusion algorithms to develop a robust visual question-answering system. The integration of visual and textual modalities enables a comprehensive understanding of the image content and facilitates the generation of accurate and meaningful questions and answers.

Chapter 2

Literature Survey

2.1 Medical Visual Question Answering: A Survey

Medical Visual Question Answering (VQA), in combination with medical artificial intelligence bring forward the challenges of VQA in real life. The goal of medical VQA is to predict plausible and convincing answers to clinically relevant questions based on medical images and natural language input.

[Lin et al., 2021] highlights the need for specific investigation and exploration of medical VQA due to its unique task features. It focuses on publicly available medical VQA datasets, the authors provide a comprehensive analysis of the data sources, data quantity, and task features. The paper highlights the importance of dataset size and diversity in training robust medical VQA models while general-domain VQA has been extensively studied, the medical VQA domain requires specific attention due to the intricacies and complexities associated with medical data. Furthermore, the paper addresses issues related to medical terminology, understanding complex medical images, and interpreting context-dependent questions in a clinical context. The paper emphasizes the need for domain knowledge and expertise in generating accurate answers and highlights the potential impact of medical VQA in clinical decision-making and healthcare applications.

Overall, the paper serves as a comprehensive survey of the field of medical VQA. It covers the publicly available medical VQA datasets, reviews the existing approaches and methodologies, and analyzes the specific challenges associated with medical VQA. The work provides valuable insights and lays the foundation for future research and advancements in the field of medical VQA.

2.2 Multiple Meta-model Quantifying for Medical Visual Question Answering

Multiple Meta-model addresses the problem of effectively utilizing meta-data within the dataset for transfer learning in the field of medical Visual Question Answering (VQA). [Do et al., 2021] presents a novel approach called Multiple Meta-model Quantifying (MMQ) that aims to overcome these limitations and improve the performance of medical VQA systems.

The proposed MMQ method is designed to increase the amount of meta-data available for training by leveraging auto-annotation techniques. By automatically annotating the data, the method effectively expands the meta-data, enabling better feature extraction and transfer learning. This is particularly useful in medical VQA, where obtaining labeled data can be challenging and time-consuming. The authors also address the issue of noisy labels that may occur during the auto-annotation process. They leverage the uncertainty of predicted scores during the meta-agnostic process to handle noisy labels and improve the robustness of the meta-models.

The output of the MMQ method is a set of meta-models that contain robust features specifically tailored for medical VQA tasks. These meta-models capture the important visual and semantic information from the medical images, enabling accurate prediction of answers to clinically relevant questions. The authors compare the performance of their approach with state-of-the-art methods on two challenging medical VQA datasets. The experimental results demonstrate that the MMQ method achieves superior accuracy compared to existing approaches, while not requiring the use of external data for training the meta-models.

Overall, the paper introduces the MMQ method as a novel approach for improving the performance of medical VQA systems by effectively utilizing meta-data within the dataset. By leveraging auto-annotation techniques and addressing noisy labels, the method enhances the transfer learning process and generates robust meta-models. The experimental results validate the effectiveness of the proposed approach, showcasing its potential to advance the field of medical VQA.

2.3 Sequential VQA with Attention for Medical Visual Question Answering

Sequential VQA focuses on the approach to the Medical Visual Question Answering (VQA-Med 2020) task of the ImageCLEF 2020 competition. The goal of the task is to generate answers to questions based on both an image and the question itself. The authors propose an encoder-decoder architecture to accomplish this task.

The encoder-decoder architecture consists of an encoder module and a decoder module. The encoder module takes two inputs: the image feature vectors obtained from the VGG16 model, and a vector representation for each question using the BERT language model. The question features are further processed using the self-attention mechanism to obtain attention features. These question attention features, along with the image features, are fused using a multi-modal factorized bilinear pooling technique called MFB. This fusion allows the model to capture the relationships between image and question features.

The fused features are then subjected to self-attention to obtain fused attention features. The thought vectors are obtained by concatenating these fused attention features with the hidden states of the encoder LSTM. The decoder module generates the answer word by word based on the input question and image.

The performance of the proposed model is evaluated in terms of accuracy and BLEU score. The best-performing model achieved an accuracy of 37.8% and a BLEU score of 0.439. The results demonstrate the effectiveness of the model in generating accurate answers for medical visual question-answering tasks.

[Verma and Ramachandran, 2020] highlights the importance of VQA in the healthcare sector, where it can aid in the interpretation of radiology images and support clinical decision-making. It also discusses the advancements in artificial intelligence, such as BERT language models and fusion techniques based on bilinear pooling and attention mechanisms, which have contributed to the progress in VQA research.

The paper concludes by discussing related works in the field of VQA-Med and provides an overview of the dataset used for the task. It describes the model architecture, the evaluation process, and the achieved results. Finally, the authors present their conclusions and highlight potential directions for future work in the field of medical visual question answering.

2.4 Overview of the Medical Visual Question Answering Task

[Abacha et al., 2019] provides an overview of the VQA-Med task conducted as part of the ImageCLEF 2019 competition. The task aimed to answer medical questions based on the visual content of radiology images. The authors focused on four categories of clinical questions: Modality, Plane, Organ System, and Abnormality, which varied in difficulty and required different approaches such as classification and text generation.

To ensure that all questions could be answered solely based on the image content without additional medical knowledge, the authors created a new dataset consisting of 4,200 radiology images and 15,292 question-answer pairs. The challenge received participation from 17 teams, who employed various techniques including transfer learning, multi-task learning, and ensemble methods.

The best-performing team achieved a BLEU score of 64.4% and an accuracy of 62.4%. The results indicate the potential of VQA in medical image interpretation and clinical decision support.

The paper emphasizes the opportunities offered by recent advances in artificial intelligence for clinical decision support and the automatic interpretation of medical images. They highlight the significance of systems capable of understanding clinical images and answering questions related to their content in clinical education, decision-making, and patient education. domain.

In VQA-Med 2019, the authors curated a larger dataset of radiology images and medical questions that required answers based solely on the image content and focused on categories such as Modality, Plane, Organ System, and Abnormality. The difficulty levels varied, with some categories lending themselves to classification tasks while others required answer generation.

The paper provides detailed information on the task description, dataset creation process, and the VQA-Med-2019 dataset itself. It also presents the evaluation methodology and discusses the results obtained in the challenge.

Overall, the paper presents an overview of the VQA-Med task, highlights the challenges specific to medical visual question answering, and provides insights into the dataset, evaluation, and results obtained during the competition.

2.5 SF-YOLOv5: A Lightweight Small Object Detection Algorithm Based on Improved Feature Fusion Mode

SF-YOLOv5 presents a novel in the domain of object detection, designed specifically for detecting small objects in images. [Liu et al., 2022] address the challenge of accurately detecting small objects, which is often challenging due to their limited visual details and low resolution.

The SF-YOLOv5 algorithm builds upon the YOLOv5 architecture, which is a popular object detection model known for its speed and accuracy. The authors propose improvements to the feature fusion mechanism of YOLOv5 to enhance the detection performance for small objects. By optimizing the feature fusion process, SF-YOLOv5 aims to improve the model's ability to capture and represent fine-grained details necessary for accurate small object detection.

To evaluate the performance of SF-YOLOv5, the authors conduct extensive experiments on benchmark datasets commonly used in object detection tasks. They compare SF-YOLOv5 with other state-of-the-art object detection algorithms, including YOLOv4, EfficientDet, and FCOS. The evaluation metrics include standard object detection metrics such as precision, recall, and mean average precision (mAP).

The experimental results demonstrate that SF-YOLOv5 achieves competitive performance in terms of object detection accuracy while maintaining a lightweight and efficient design. It outperforms existing algorithms in detecting small objects and demonstrates improved localization accuracy.

The article also discusses the practical applications and potential use cases of SF-YOLOv5, particularly in scenarios where small object detection is crucial, such as surveillance systems, autonomous vehicles, and robotics.

In conclusion, the journal article presents SF-YOLOv5, a lightweight small object detection algorithm based on an improved feature fusion mode. The algorithm demonstrates promising results in accurately detecting small objects while maintaining computational efficiency. The research contributes to the field of object detection by addressing the specific challenges associated with small object detection and providing a practical solution for applications that require the precise detection of such objects.

2.6 ChatGPT versus Traditional Question Answering for Knowledge Graphs

ChatGPT explores the characteristics and limitations of conversational AI and question-answering systems (QAS) for knowledge graphs (KGs). Conversational AI aims to simulate human-like conversations, while QAS retrieves information from KGs by translating natural language questions into structured queries.

[Omar et al., 2023] presents a comprehensive study comparing two representative conversational models, ChatGPT and Galactica, against KGQAN, a state-of-the-art QAS. The evaluation is conducted on four real KGs from different application domains to identify the limitations of each system category. The goal is to identify research opportunities to enhance QAS with chatbot capabilities for KGs.

The paper highlights the growth of conversational AI models and the potential for combining their capabilities with structured KG information to provide more accurate answers in specific domains. Future KG chatbot systems would require a QAS mechanism to update responses with the most recent information retrieved from a given KG.

KGQAN is identified as a state-of-the-art QAS that can answer questions from arbitrary KGs across various application domains. It formalizes question-answering as a three-fold problem: question understanding, linking, and filtering of final answers. KGQAN utilizes pre-trained language models such as BART and GPT-3 for question understanding and employs a just-in-time approach based on word embedding models like FastText for linking and filtering.

The paper proposes a comparative framework to systematically review the capabilities of conversational AI language models versus KGQASs in answering questions. The framework is used to evaluate state-of-the-art language models like ChatGPT and Galactica against KGQAN. The evaluation is performed on real KGs from diverse domains, and the current limitations of each category are identified.

Furthermore, the paper presents open research challenges and opportunities for developing a KG chatbot that incorporates the merits of conversational AI models and QASs.

In summary, the paper provides an in-depth analysis of conversational AI and traditional QAS approaches for KGs. It compares different models, evaluates their performance on real KGs, and suggests future research directions for creating advanced KG chatbot systems.

Chapter 3

Proposed Methodology

3.1 Problem Statement

The proposed methodology combines YOLO-based object detection with a BERT-based question-answering system to enable accurate question answering based on user input. The problem statement revolves around extracting image features, assigning class labels to the objects detected in the image, and generating appropriate answers to questions posed by the user [3.1].

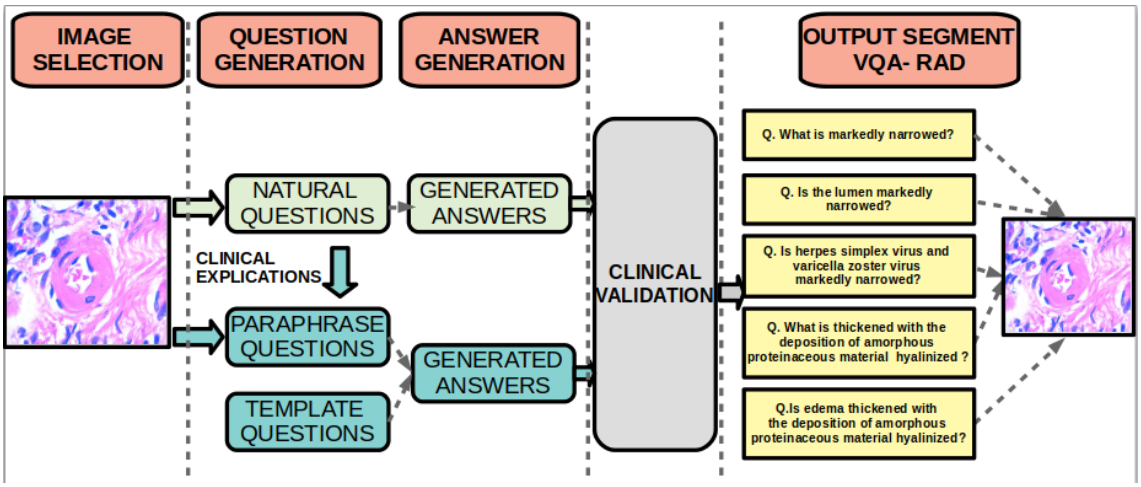

Figure 3.1: Block Diagram of VQA system for medical image data

1. Problem Statement: The goal is to develop a robust question-answering system that can effectively analyze images using YOLO-based object detection and generate accurate answers based on the detected objects and user questions. The system should be able to handle a wide range of questions and provide meaningful responses that are relevant to the content of the image.

2. YOLOv5-based Object Detection: The first part of the methodology involves utilizing the YOLO (You Only Look Once) [Jiang et al., 2022] algorithm for object detection in images. YOLO is a real-time object detection framework that can identify multiple objects within an image and provide their bounding box coordinates [3.2]. By applying YOLO, the system extracts the image features and identifies the objects present, along with their respective class labels.

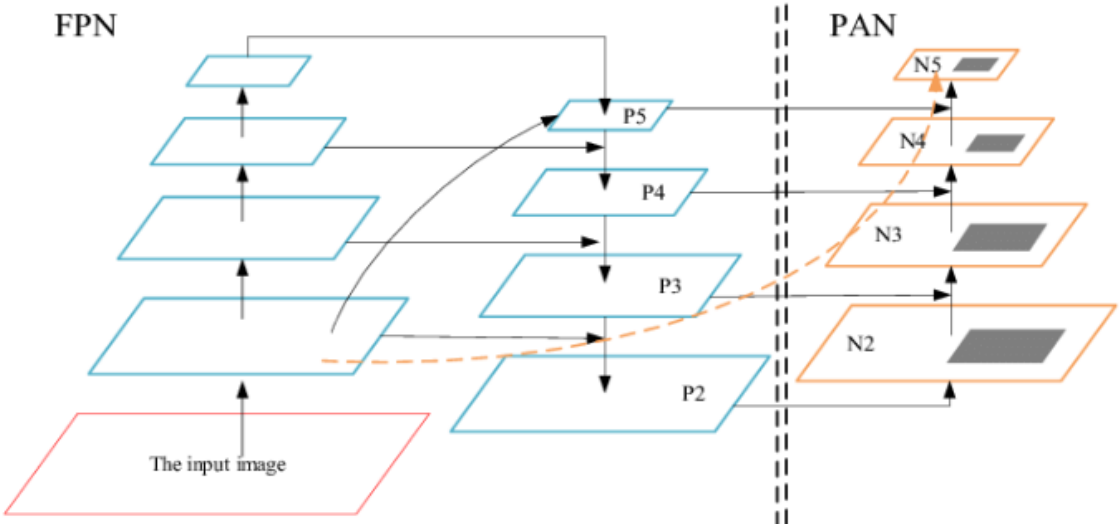

Figure 3.2: Baseline Architecture of Yolo V5 model

3. BERT-based Question Answering: The next step is to utilize the BERT (Bidirectional Encoder Representations from Transformers) model for question answering. BERT is a powerful natural language processing (NLP) model that has achieved state-of-the-art performance in various NLP tasks. In this methodology, BERT is employed to understand the user's question and generate appropriate answers based on the image features obtained from the previous step.

4. Integration of YOLO and BERT: The extracted image features and object class labels obtained from YOLO are combined with the user's question input and fed into the BERT model. The BERT model takes the concatenated inputs and performs fine-tuning to learn the relationship between the image features, class labels, and the corresponding questions. By training the BERT model on a labeled dataset, it becomes capable of predicting accurate answers to new user questions.

5. Prediction and Answer Generation: Once the BERT model is trained, it can be used to predict answers for user questions based on the image

features and class labels. The trained model takes a user's question as input, along with the corresponding image features and class labels. It then generates the most appropriate answer by leveraging the learned relationships captured during training.

6. Evaluation and Validation: The proposed methodology should be evaluated using appropriate metrics to assess the accuracy and performance of the question-answering system. Evaluation can involve comparing the generated answers with ground truth answers or conducting user studies to assess the system's effectiveness in providing relevant and meaningful responses.

   By combining YOLO-based object detection for image feature extraction as shown in [3.4] and BERT-based question answering for generating answers, the proposed methodology aims to address the problem of accurately answering user questions based on image content. The integration of computer vision and NLP techniques enables a comprehensive understanding of both visual and textual information, leading to more precise and context-aware answers as shown in [3.3].

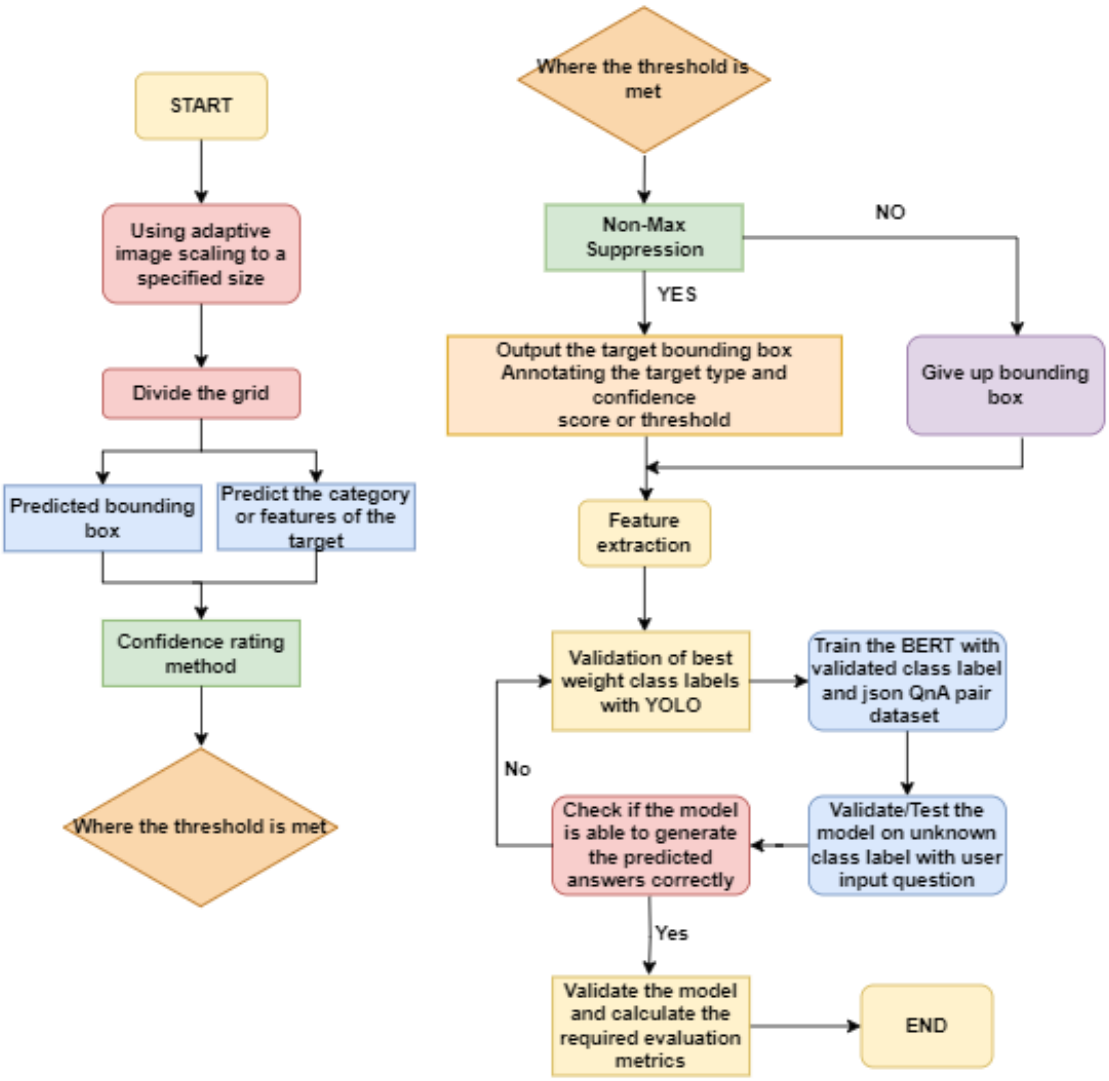

Figure 3.3: Workflow of our model

3.1.1 Algorithm

Algorithm 1 Train and Evaluate a Question Answering NLP Model

procedure TRAINANDEVALUATEQAMODEL($model, trainDataloader$,)
$validationDataloader, optimizer, epochs$)
Initialize variables
Set number of epochs to $epochs$
Initialize lists to store training losses, validation accuracies, and learning rates
for epoch in range(1, epochs + 1) **do**
▷ Training loop
Set model to train mode
Set training loss to 0
Set number of training steps to 0
for step, batch in enumerate(trainDataloader) **do**
Convert batch to PyTorch tensors
Set model inputs
Zero out the gradients
Get model outputs
Compute loss
Backpropagate the loss
Update the model parameters
Add the loss to the training loss
Increment the number of training steps
end for
▷ Compute training loss
Divide the training loss by the number of training steps
Add the training loss to the list of training losses
▷ Validation loop
Set model to evaluation mode
Set number of correct answers to 0
Get the list of valid examples
for batch in validationDataloader **do**
Convert batch to PyTorch tensors
Get model outputs
Get predicted start and end positions
Get the predicted answer
Check if the predicted answer is correct
Increment the number of correct answers
end for
▷ Compute validation accuracy
Divide the number of correct answers by the number of valid examples
Add the validation accuracy to the list of validation accuracies
▷ Store learning rate
Get the learning rate from the optimizer
Add the learning rate to the list of learning rates
Print the training loss
Print the validation accuracy
end for
end procedure

3.2 Data Collection

In the data collection phase, you have followed a specific process to gather the raw data and create an annotated dataset for your YOLO-based object detection model. The detailed steps involved in this process are as follows:

1. Selection of Source Material: You mentioned that you collected raw data from the Pranab Dey book. This implies that you have chosen a specific book or resource as the primary source for your dataset. It could be a book that contains images relevant to the objects or classes you want to detect.

2. Image Extraction: Once you have identified the relevant images from the source material, you extract these images to form the initial dataset. This step involves manually extracting the images from the book or resource, ensuring that you obtain high-quality images suitable for annotation.

3. Manual Annotations: After extracting the images, you performed manual annotations on the dataset. Manual annotation involves marking the objects or regions of interest within the images and assigning class labels to them. In your case, you used the YOLO workspace provided by Roboflow, which likely offers tools and interfaces to annotate the images and define bounding boxes around the objects.

4. Data Augmentation: To enhance the diversity and variability of your dataset, you applied data augmentation techniques. Data augmentation involves applying various transformations or modifications to the original images, such as rotations, translations, scaling, flips, and color manipulations. These transformations help to increase the robustness and generalization capability of your object detection model.

5. Class Labeling: As part of the manual annotation process, you assigned class labels to the annotated objects within the images. Class labels represent the categories or types of objects you want your model to detect. For each annotated object, you manually specified the corresponding class label, ensuring that it accurately represents the object's identity or category.

   By following these steps, you have collected a dataset with annotated images, bounding box annotations, and assigned class labels. This dataset serves as the foundation for training and evaluating your

YOLO-based object detection model. The manual effort involved in data collection and annotation is crucial to ensure the quality and accuracy of the dataset, which ultimately influences the performance of your model.

3.2.1 Data Preprocessing

Data preprocessing in YOLO (You Only Look Once) involves several steps to prepare the dataset for training an object detection model. The main objective of data preprocessing is to standardize and format the data in a way that is compatible with the YOLO architecture and optimize the training process. Here is an overview of the typical data preprocessing steps in YOLO, along with the reasons for performing each step:

1. Image Resizing: In YOLO, images are usually resized to a fixed input size, 480 X 640, that the model expects. Resizing ensures that all images have the same dimensions, which is necessary for efficient batch processing during training. Additionally, resizing helps to maintain a consistent aspect ratio across the dataset, preventing distortion or loss of information.

2. Anchor Box Generation: YOLO utilizes anchor boxes to predict object bounding boxes at different scales and aspect ratios. The anchor boxes are pre-defined boxes that represent different object sizes and shapes. During data preprocessing, anchor boxes are typically generated by analyzing the dataset to determine the clusters of object sizes and aspect ratios. These anchor boxes are then used during training to guide the model's predictions.

3. Label Encoding: In YOLO, each object bounding box annotation is associated with a class label. Label encoding is performed to convert the class labels into numerical representations that the model can understand. This usually involves assigning a unique numerical identifier to each class label. For example, if you have three classes (person, car, and dog), you may assign the numerical labels 0, 1, and 2, respectively.

4. Data Augmentation: Data augmentation techniques are applied to artificially increase the diversity and variability of the dataset. Augmentation helps the model generalize better by introducing variations in image appearance, such as rotations, translations, flips, and color

changes. This helps the model learn to detect objects under different conditions, improving its robustness and performance on unseen data.

5. Bounding Box Encoding: The annotated bounding boxes for objects in the images need to be encoded in a specific format that the model can interpret during training. YOLO commonly uses the ”YOLO format” or ”darknet format” for bounding box encoding. This format includes the normalized coordinates of the bounding box's center, width, height, and class label associated with the object.

6. Splitting into Training and Validation Sets: To evaluate the performance of the trained model, it is essential to have a separate validation set. The dataset is typically split into training and validation subsets, with a certain percentage of images reserved for validation. This allows monitoring the model's performance on unseen data and helps prevent overfitting.

   Overall, data preprocessing in YOLO ensures that the dataset is appropriately formatted, standardized, and augmented to improve the model's ability to detect objects accurately and generalize well to unseen data. Each step has its specific purpose, contributing to the overall effectiveness and performance of the trained YOLO model.

3.3 Approach

3.3.1 YOLO-based VQA model

The architecture you described combines YOLO (You Only Look Once) for object detection and localization with a CNN-based VQA model. Here is a detailed explanation of the deep architecture:

1. YOLO for Object Detection and Localization: YOLO is a popular object detection algorithm that operates in a single pass over the entire image. It divides the image into a grid and predicts bounding boxes and class probabilities for each grid cell. The YOLO architecture consists of several convolutional layers followed by fully connected layers. The main components of the YOLO architecture include:

   - Input Layer: The input to the YOLO model is an image. - Convolutional Layers: The convolutional layers capture various image features at different spatial scales. - Anchor Boxes: YOLO uses anchor boxes of different sizes and aspect ratios to predict bounding boxes

for objects. - Bounding Box Regression: The YOLO model predicts bounding box coordinates relative to each anchor box. - Class Prediction: The model predicts class probabilities for each object category present in the image. - Non-maximum Suppression: After obtaining the bounding box predictions, non-maximum suppression is applied to remove duplicate or overlapping bounding boxes. - Output: The final output of the YOLO model consists of the predicted bounding boxes and their associated class probabilities.

2. CNN for Visual Question Answering (VQA): After extracting the image features and localizing objects using YOLO, the architecture incorporates a CNN-based VQA model to answer questions based on the image and textual input. Here is a general overview of the CNN-based VQA model:

   - Image Encoding: The image features extracted by YOLO are encoded using a CNN. The CNN processes the image regions within the bounding boxes to capture high-level visual representations. - Question Encoding: The textual input (question) is encoded using a word embedding technique, such as Word2Vec or GloVe, to convert words into dense numerical vectors. - Fusion: The encoded image features and question embeddings are combined (fused) to create a joint representation that captures the relationship between the image and the question. - Attention Mechanism: An attention mechanism is often employed to focus on the relevant image regions or words in the question during the fusion process. This helps the model align the visual and textual information effectively. - Prediction: The fused representation is passed through additional layers, such as fully connected layers or recurrent neural networks, to generate the final prediction or answer to the question.

   The YOLO-based CNN VQA model leverages the power of YOLO for accurate object detection and localization. The localized objects are then processed through a CNN-based VQA model that combines visual and textual information to generate answers to user questions.

3.3.2 YOLO Model Summary

YOLOv5 is an object detection algorithm that builds upon the success of the YOLO (You Only Look Once) [Jiang et al., 2022] family of models. It is a lightweight and efficient model that achieves high accuracy in real-

time object detection tasks. Here is a summary of the YOLOv5 model architecture as shown in [3.5]:

1. Backbone: YOLOv5 utilizes a deep neural network backbone to extract features from the input image. The backbone is typically a variant of the EfficientNet architecture, which is a highly efficient and powerful convolutional neural network (CNN). The backbone network consists of multiple convolutional layers,including both standard convolutional layers and more advanced ones like depthwise separable convolutions that capture increasingly complex image features. These layers capture increasingly complex visual patterns and encode them into feature maps.

2. Neck: YOLOv5 introduces a "neck" component to further enhance the feature representation. The neck is responsible for fusing and aggregating features from different scales in the backbone network. It enhances the representation power of the backbone features and enables the model to detect objects of various sizes. YOLOv5 uses techniques like feature pyramid networks (FPN) or spatial pyramid pooling (SPP) to achieve multi-scale feature fusion.This multi-scale feature fusion allows YOLOv5 to detect objects of various sizes and maintain a high level of accuracy.

3. Head: The head of the YOLOv5 model performs the final detection and classification. It consists of a series of convolutional layers that predict bounding box coordinates and object class probabilities.These layers predict bounding box coordinates (x, y, width, height) for object localization and class probabilities for object classification.YOLOv5 employs anchor boxes of different sizes and aspect ratios to handle objects with varying scales. The predictions are made across different grid cells at different scales, enabling the model to detect objects throughout the image.

4. Loss Function: During training, YOLOv5 model [Jocher et al., 2020] Ultralytics have used Binary Cross-Entropy with Logits Loss function from PyTorch for loss calculation of class probability and object score, which encompasses both localization loss (to ensure accurate bounding box predictions) and classification loss (to accurately predict object classes).These losses are computed based on the predicted bounding boxes and the ground truth annotations.These losses are computed based on the predicted bounding boxes and the ground

truth annotations.

$$Hp(q) = -\frac{1}{N}\sum_{i=1}^{N} yi.\log(P(yi)) + (1-yi).\log(1-p(yi))$$

5. Inference: During inference, YOLOv5 processes the input image and passes it through the backbone and neck to extract features. The head then performs object detection by predicting bounding box coordinates and class probabilities. Non-maximum suppression is applied to remove duplicate or overlapping bounding boxes, resulting in the final object detection output.

   YOLOv5 is known for its efficiency, achieving real-time object detection performance on both CPUs and GPUs. It provides a good balance between accuracy and speed, making it suitable for a wide range of applications, including real-time object detection in videos, autonomous driving, surveillance, and robotics.

   It's worth noting that YOLOv5 is an evolution of the original YOLO models, with improvements in architecture, performance, and efficiency. The specific details of the YOLOv5 model, such as the number of layers, filters, and input size, may vary depending on the specific configuration used during training.

3.3.3 BERT-based NLP Model Summary

In this paper, we explain the detailed summary of the BERT-based NLP model [Liu et al., 2019] in the context of generating Visual Question Answers (VQA) using the outputs of object detection with YOLOv5:

1. Input Encoding: The BERT-based NLP model takes as input a question related to the objects detected in the image. The question is tokenized, where each token represents a word or subword unit. The tokens are then converted into their corresponding embeddings.

2. Pretrained BERT Model: The BERT model consists of a transformer-based architecture that has been pretrained on a large corpus of text data. It leverages the bidirectional nature of the transformer to capture the contextual relationships between words. The pretrained BERT model is used as a feature extractor in this architecture.

3. Question Encoding: The tokenized question is fed into the BERT model, which processes it through several transformer layers. These

layers capture the contextual information by attending to the surrounding words in the question. The output of the BERT model for the question is a sequence of contextualized embeddings that represent the encoded question.

4. Region of Interest (ROI) Features: Simultaneously, the ROI extracted from the image using YOLOv5 is processed through the object detection pipeline, including the backbone and neck components. This results in a set of high-level visual features that represent the detected objects within the ROI.

5. Fusion of Visual and Textual Features: The encoded question embeddings and the visual features from the ROI are fused together to create a joint representation of the visual and textual information. This fusion can be achieved through various techniques, such as concatenation, element-wise multiplication, or attention mechanisms. The goal is to combine the complementary information from both modalities.

   6. Additional Layers: The fused features are passed through additional layers, such as fully connected layers or attention mechanisms, to further process and refine the joint representation. These layers aim to capture the relationships between the visual and textual features and learn the associations between objects and the corresponding textual information.

   7. Answer Generation: The final output of the model is a classification head that predicts the answer to the question based on the fused features. The classification head can be a fully connected layer followed by softmax activation, which generates a probability distribution over possible answers. The answer with the highest probability is considered to be the predicted answer.

   By combining the power of BERT's language understanding capabilities with the visual features extracted from the ROIs using YOLOv5 [Lohith et al., 2023], this architecture enables the model to generate accurate and contextually relevant answers to questions about the detected objects in an image. The joint representation of visual and textual features allows the model to leverage both modalities and effectively reason about the relationships between objects and the corresponding textual context.

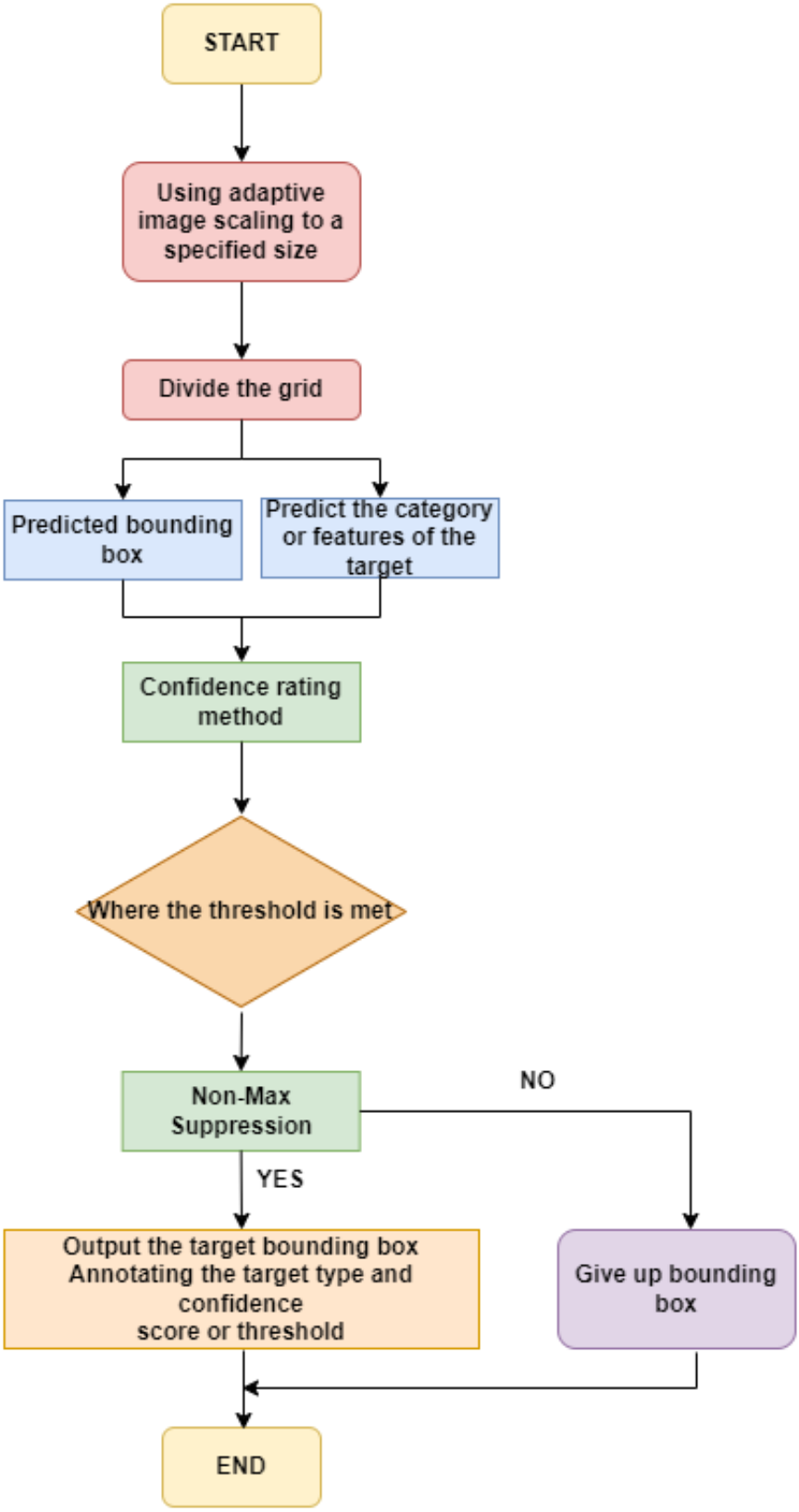

Figure 3.4: Flow diagram of YOLOv5 model for object localization

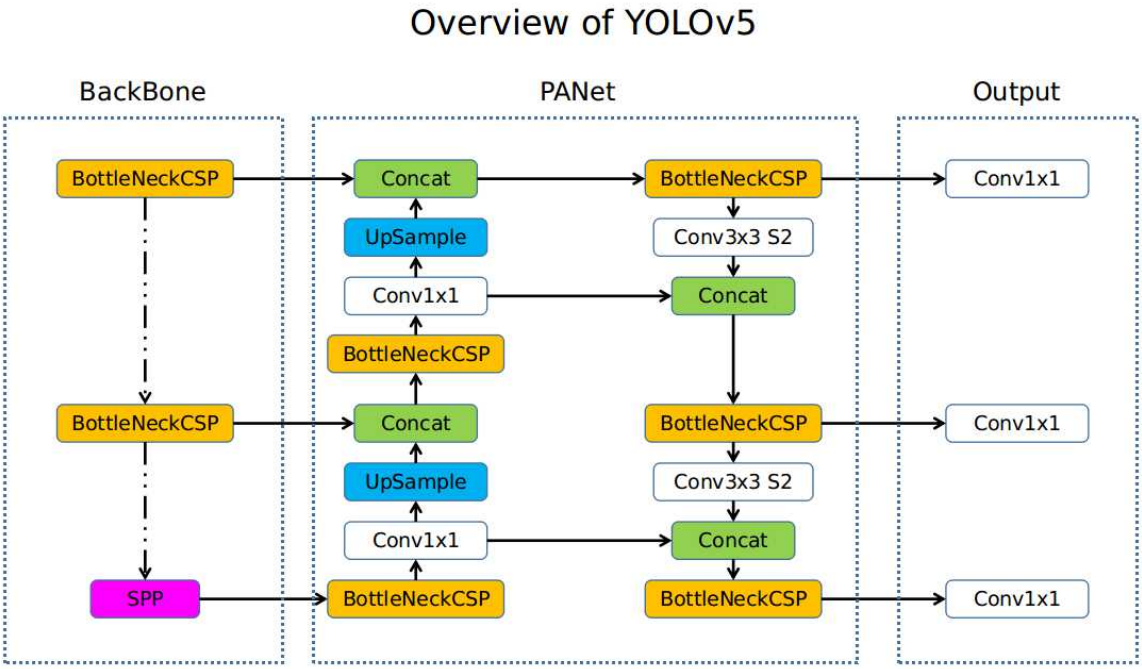

Figure 3.5: YOLOv5 model architecture

Chapter 4

Result and Analysis

In this work, the procedure used for building a Visual Question Answering model are described in the section below.

1. Dataset Selection: The project requires a dataset that includes both images and corresponding questions about the objects in the images. A suitable dataset for this task is chosen, such as VQA (Visual Question Answering) datasets that provides images, questions, and their corresponding answers.

2. Data Preprocessing: The selected dataset undergoes preprocessing steps to prepare it for training the YOLOv5 object detection model and the BERT-based NLP model. This includes tasks such as resizing images to a consistent resolution, splitting the dataset into training, validation, and testing sets, and encoding questions into a suitable format for BERT input.

3. YOLOv5 Training: The YOLOv5 model is trained on the training set of the dataset using the annotated bounding box information to detect objects within the images. The model is trained with appropriate hyperparameters, loss functions, and optimization techniques to optimize the object detection task. The training process may involve multiple epochs to improve the model's performance.

4. Object Detection Evaluation: Once the YOLOv5 model is trained, it is evaluated on the validation set or a separate evaluation set to assess its performance in detecting objects accurately. Evaluation metrics such as mean Average Precision (mAP) and Intersection over Union (IoU) are calculated to measure the quality of object detection results.

5. BERT-based NLP Model Training: Next, the BERT-based NLP model is trained using the preprocessed dataset. The model is trained to gen-

erate answers to the questions based on the visual features extracted from the ROIs using YOLOv5. The BERT model's parameters are fine-tuned using the training set, and appropriate optimization techniques and loss functions are used for training.

6. Joint Training and Fine-tuning: In some cases, the YOLOv5 and BERT models can be jointly trained or fine-tuned together to optimize the overall performance of the system. This step involves training the models simultaneously, considering both the object detection and question-answering objectives.

7. Evaluation and Performance Metrics: The trained BERT-based NLP model is evaluated on the testing set or a separate evaluation set to assess its performance in generating accurate answers to the questions based on the detected objects. Evaluation metrics such as accuracy, BLEU score, ROUGE score, or other relevant metrics are calculated to measure the quality of the generated answers.

8. Comparison and Analysis: The performance of the YOLOv5 object detection model and the BERT-based NLP model is analyzed individually, as well as in combination, to understand their strengths, weaknesses, and overall effectiveness in the task of generating visual question answers. The results are compared with baseline models or previous works to evaluate the improvement achieved by the proposed architecture.

9. Further Experimentation: Based on the analysis of the experimental results, additional experiments or modifications to the architecture can be performed to enhance the performance of the system. This may include fine-tuning hyperparameters, exploring different fusion techniques for combining visual and textual features or incorporating additional components into the architecture.

   By following this experimental setup, the project aims to evaluate the effectiveness of the YOLOv5-based object detection and BERT-based NLP models for generating accurate and contextually relevant question answers based on the detected objects in images. The experiments help assess the performance of the models individually and in combination, providing insights into their strengths and areas for improvement

Comparison of Original and Detected Images

- **Image 1:**

Figure 4.1: Oval parabasal cells, renal tubules and squamous cells detected in image

- **Image 2:**

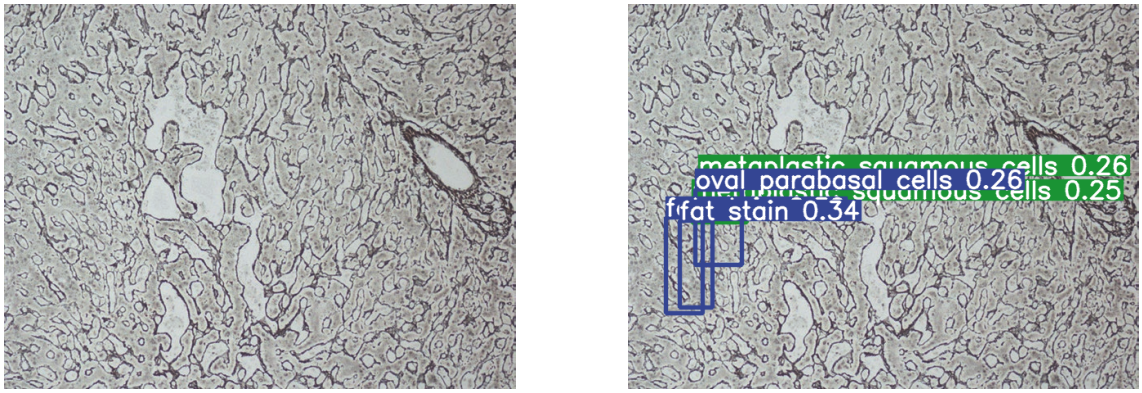

Figure 4.2: Fat stain, oval parabasal cells and squamous cells detected in image

- **Image 3:**

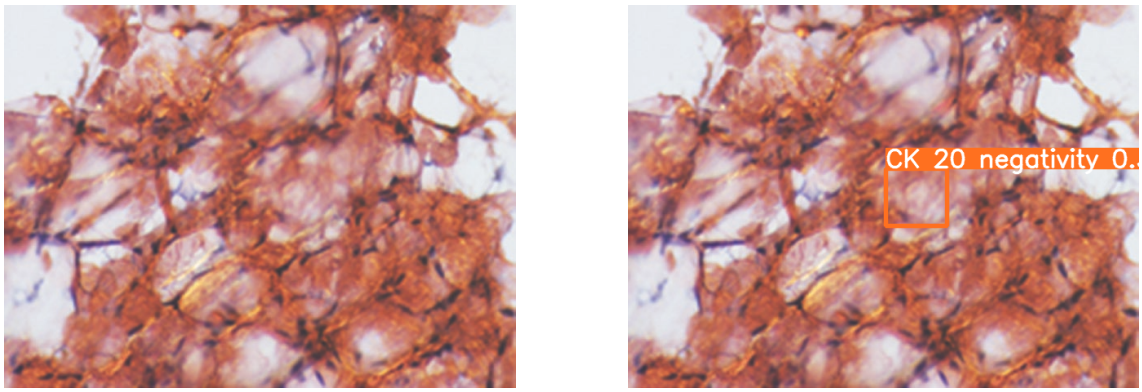

Figure 4.3: CK 20 negativity detected in image

- **Image 4:**

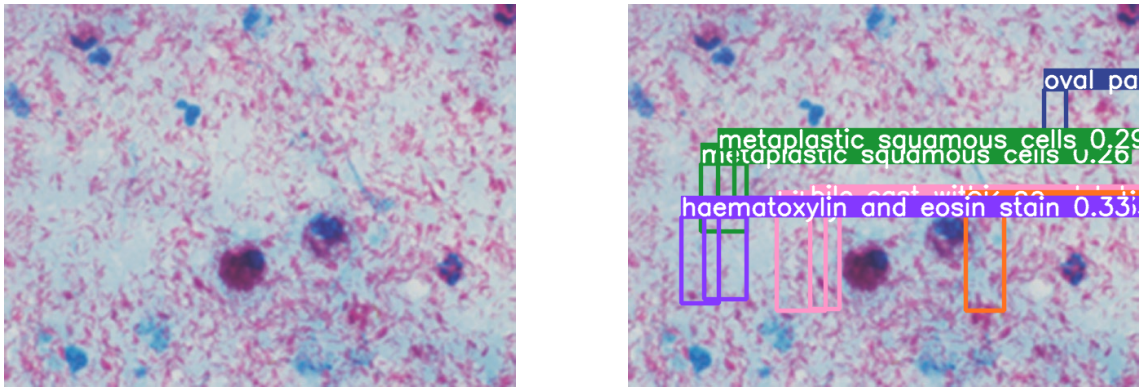

Figure 4.4: Oval parabasal cells, squamous cells, renal tubules and hematoxylin and eosin stain detected in image

```
Unique Class Labels:
---------------------
squamous cells
haemosiderin pigment
Question-Answer Pairs:
----------------------
Class Label: haemosiderin pigment
----------------------
Question:  What is the purpose of the Perls' reaction in cytology smears?
Answer: The purpose of the Perls' reaction is to detect the presence of iron in the form of
haemosiderin pigments in the cytology smear. It is commonly used in the diagnosis of diseases
such as hemosiderosis
----XXX----
Question:  What does the Perls' reaction show in the cytology smear?
Answer: The Perls' reaction shows dark blue haemosiderin pigments in the cytology smear
----XXX----
Class Label: squamous cells
----------------------
Question: How are metaplastic squamous cells diagnosed?
Answer: Diagnosis of metaplastic squamous cells typically involves microscopic examination of
tissue samples obtained through a biopsy or cytology. Pathologists can identify the presence
of metaplastic squamous cells and differentiate them from normal cells based on their
morphological characteristics and staining patterns
----XXX----
Question: How are squamous cells related to acute inflammation?
Answer: Squamous cells can be involved in acute inflammation as part of the body's immune response.
During acute inflammation, the affected area may undergo changes such as increased blood flow and
increased permeability of blood vessels. This can lead to the migration of immune cells, including
squamous cells, to the site of inflammation to help combat infection or injury
----XXX----
Question: What are squamous cells?
Answer: Squamous cells are flat, thin cells that form the epithelial lining of various organs and
structures in the body. They are characterized by their flattened shape and tightly packed arrangement
----XXX----
```

Figure 4.5: Training samples for our VQA dataset

The diagram [4.5] illustrates the training process of a VQA model using extracted class labels and corresponding question-answering pairs. At the time of training, the VQA model associated with question-answer pair and a class label containing relevant visual information. The model's objective is to learn to understand the question and generate accurate answer based on the image features.
The figure [4.6] depicts the testing phase of a VQA model using extracted class labels and their corresponding keywords. This testing phase involves generating natural language-based answers for user input questions. The VQA model utilizes these extracted class labels as a form of contextual understanding of the user-input question. By matching the keywords in the question with the relevant class labels, then using this contextual understanding, the VQA model generates a natural language-based answer in response to the user's question. The answer is formulated based on the combined knowledge gained from the training data, including the relationship between the extracted class labels and the corresponding questions and answers.

```
Context:
The image displays a histological sample showing an irregular tear in the tissue. Multiple air bubbles have ruptured the
tissue. The air bubbles are produced due to uncontrolled temperature in the water bath. Note the characteristic multiple
irregular tear of the tissue. Tissue tears of this nature often result from mechanical forces,such as trauma, surgical
manipulation, or external injuries. The irregularity of the tear suggests a non-linear or jagged disruption in the tissue
structure. Tear in the tissue appears due to uneven cutting edge of the knife. Histological examination plays a crucial role
in assessing the extent of tissue damage and identifying the underlying causes. Further investigation, including a detailed
clinical history, additional imaging studies, or consultation with relevant specialists, may be required to determine the
specific cause and potential implications of the tissue tear. It is important to note that without specific information about
the location, patient history, or additional clinical context, providing a more precise analysis or explanation of the tissue
tear seen in the picture is challenging. Consultation with a qualified healthcare professional or a specialist in pathology
or surgery would be advisable to obtain a comprehensive evaluation and interpretation of the findings. Multiple air bubbles
produced due to uncontrolled temperature in the water bath caused the characteristic multiple irregular tear of the tissue.
The picture reveals an irregular tear in the tissue. The tissue tear could be caused by trauma, surgical manipulation, or
external injuries. The tissue tear appears irregular and non-linear, suggesting a jagged disruption in the tissue structure.
Further investigations such as a detailed clinical history, additional imaging studies, or consultation with relevant
specialists may be required to determine the specific cause and implications of the tissue tear. Factors such as the location
of the tear, surrounding structures, and the patient's clinical presentation should be considered for accurate diagnosis and
appropriate management. Treatment options may vary depending on the severity and location of the tissue tear and may include
surgical repair, wound management, or other interventions tailored to the individual case. A qualified healthcare
professional or a specialist in pathology or surgery would be able to provide a comprehensive evaluation and interpretation
of the tissue tear findings.
--------------------------
Enter your question (or 'exit' to quit): What is the image display?
Answer: the image displays a histological sample showing an irregular tear in the tissue.
--------------------------
Enter your question (or 'exit' to quit): What have ruptured the tissue?
Answer:
--------------------------
Enter your question (or 'exit' to quit): What caused the multiple irregular tears in the tissue shown in the microphotograph?
Answer: a detailed clinical history, additional imaging studies, or consultation with relevant specialists, may be required
to determine the specific cause and potential implications of the tissue tear. it is important to note that without specific
information about the location, patient history, or additional clinical context, providing a more precise analysis or
explanation of the tissue tear seen in the picture is challenging. consultation with a qualified healthcare professional
or a specialist in pathology or surgery would be advisable to obtain a comprehensive evaluation and interpretation of the
findings.
```

Figure 4.6: Results obtained after testing VQA

4.1 Data Description

In this project, a unique dataset was curated by manually extracting images from the book ”Basic and Advanced Laboratory Techniques in Histopathology and Cytology” by Pranab Dey. These images were carefully annotated using the YOLO framework, a powerful tool for object detection, in the Roboflow workspace platform [Lin et al., 2022]. To make the dataset more diverse and robust, data augmentation techniques were applied. This involved applying various transformations, such as cropping, rotation, and flipping, to generate additional training samples. This augmentation process helps the model to generalize better and perform well on unseen data.

To train our combined YOLO and BERT-based NLP model, the dataset was split into three subsets: training, testing, and validation. The training set comprised 85% of the data, ensuring that the model learned from a substantial amount of information. The testing set, which accounted for 10% of the data, was used to evaluate the model’s performance and assess its ability to generalize to new examples. The remaining 5% constituted the validation set, which provided an independent evaluation of the model’s performance and helped fine-tune its parameters.

In parallel to the image dataset, a complementary dataset was created

specifically for question-answer pairs. This dataset facilitated the connection and correlation of the visual features extracted by the YOLO architecture with the corresponding detailed textual information. By combining the YOLO architecture for object detection and localization with the BERT architecture for natural language processing, the model was able to effectively understand and respond to questions based on the visual context of the images.

By integrating the YOLO and BERT architectures, the model achieved a powerful synergy. It successfully leveraged the visual features extracted from the images to generate accurate and contextually relevant answers to the questions posed. This integration allowed for a comprehensive understanding of the input data and enabled the model to provide insightful responses based on the combined knowledge from both the visual and textual domains.

In summary, the dataset was meticulously curated by extracting images from a specific book, annotating them using YOLO, and augmenting the data to enhance its diversity. The dataset was divided into training, testing, and validation subsets. Additionally, a separate dataset was created to pair questions with corresponding answers, enabling the model to connect visual features with textual information. The fusion of the YOLO and BERT architectures empowered the model to perform accurate object detection, localization, and question-answering, ultimately providing a comprehensive and insightful understanding of the input data.

4.2 Evaluation metric

The evaluation metric for this project can vary depending on the specific objectives and requirements. However, considering the tasks involved, such as object detection, localization, and question answering, the following evaluation metrics could be relevant:

1. Object Detection: - Average Precision (AP): This metric measures the accuracy of object detection by calculating the precision and recall of the detected objects at different IoU (Intersection over Union) thresholds. - Mean Average Precision (mAP): It is the average of AP across different object classes N, providing an overall measure of object detection performance.

$$mAP = \frac{1}{N} \sum_{i=1}^{N} (APi)$$

2. Question Answering: - Accuracy: This metric measures the percentage of correctly answered questions out of the total questions asked.

$$Accuracy = \frac{Cp}{Tp} \times 100$$

where, Cp = Number of correct predictions Tp = Total number of predictions

3. Overall Performance: - F1 Score: It is the harmonic mean of precision and recall, providing a balanced measure of model performance for tasks such as object detection and question answering. - Mean Average Precision (mAP) across both object detection and question-answering tasks: This metric combines the performance of both tasks to provide an overall evaluation of the model's effectiveness.

$$F1score = \frac{2 \times (Precision \times Recall)}{PrecisionRecall}$$

These metrics are commonly used in the respective tasks and can be adapted or supplemented with domain-specific metrics depending on the project requirements. The specific evaluation metric chosen should align with the objectives of the project and provide meaningful insights into the model's performance and capabilities.

4.3 Results

The results obtained after training and validation of YOLOv5 and BERT-base NLP model on the manually created dataset are presented in table[4.1] and table[4.2] respectively. The evaluation metrics include mean Average Precision (mAP), precision, and recall.

These results highlight the suitability of the above methods in conjunction with YOLOv5 for object detection tasks. It enables accurate and reliable object detection, making it a valuable approach in various real-world applications. Further analysis and experimentation can be conducted to explore additional performance metrics and fine-tune the model for specific use cases.

Dataset used	Methodology	mean Average Precision(mAP)	Precision	Recall
Images manually extracted from ”Basic and Advanced Laboratory Techniques in Histopathology and Cytology” by Pranab Dey.	Image segmentation + object localization using YOLOv5	0.70	0.4920	0.420

Table 4.1: Result obtained after applying YOLOv5

Dataset used	Methodology	Accuracy	Precision	Recall	F1 score
Question-Answer paired dataset manually created in **.json** format	YOLOv5 based algorithm + BERT-base-uncased NLP model	0.4667	1.00	0.50	0.67

Table 4.2: Result obtained after applying BERT-based-uncased NLP model

The table showcases the epoch results of the YOLOv5 model, along with other important parameters such as additional metrics and losses.

The graphical structure of the YOLOv5 architecture is visually represented in figure [4.7]. This architecture follows a state-of-the-art approach for object detection tasks. It is composed of several interconnected layers and components that work together to identify and localize objects within an image. The analysis of the obtained results from the YOLOv5 model is presented in the later part of this document. Each individual figure provides valuable insights into the performance and capabilities of the model. The figures showcase various aspects such as mean Average Precision (mAP), precision, recall, and other relevant metrics.

Figure 4.7: Final Result obtained from YOLOv5 graphs

Overall, the combination of the YOLOv5 architecture and the analysis of the obtained results provides valuable insights into the model's effectiveness in object detection tasks. The subsequent sections will provide a thorough exploration of each figure, enabling a comprehensive understanding of the YOLOv5 model's performance and its implications for real-world applications.

The F1 score graph, displayed in figure [4.8], provides a visual representation of the performance of the model. The F1 score is a metric that combines precision and recall to measure the overall effectiveness of the model in object detection. The F1 score curve provides a concise and visual representation of the model's performance in object detection. It helps us understand the relationship between precision, recall, and threshold values, enabling us to make informed decisions on optimizing the model for optimal object detection performance.

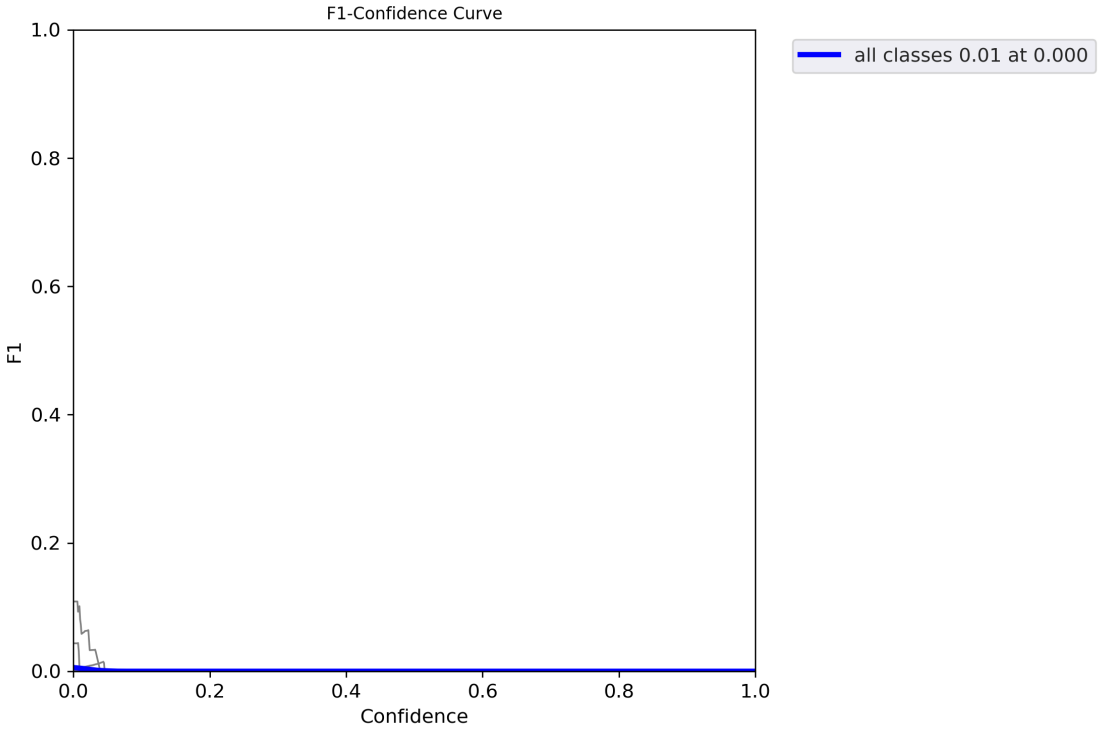

Figure 4.8: Final result of F1 curve obtained from YOLOv5

Abel's correlogram and graphically represented results of YOLOv5 are displayed in figure [4.9] and figure [4.10]. These figures provide valuable insights into the spatial distribution and patterns of the detected objects in the dataset.

The P-curve, PR-curve, and R-curve, depicted in figure [4.11], figure [4.12] and figure [4.13] respectively to provide essential information about the performance and evaluation of the object detection model. The P-curve represents the precision of the object detection model at different confidence thresholds. It illustrates how accurately the model identifies true positive detections while minimizing false positives. A higher precision value indicates a higher level of confidence in the model's predictions. The PR-curve, also known as the Precision-Recall curve, shows the trade-off

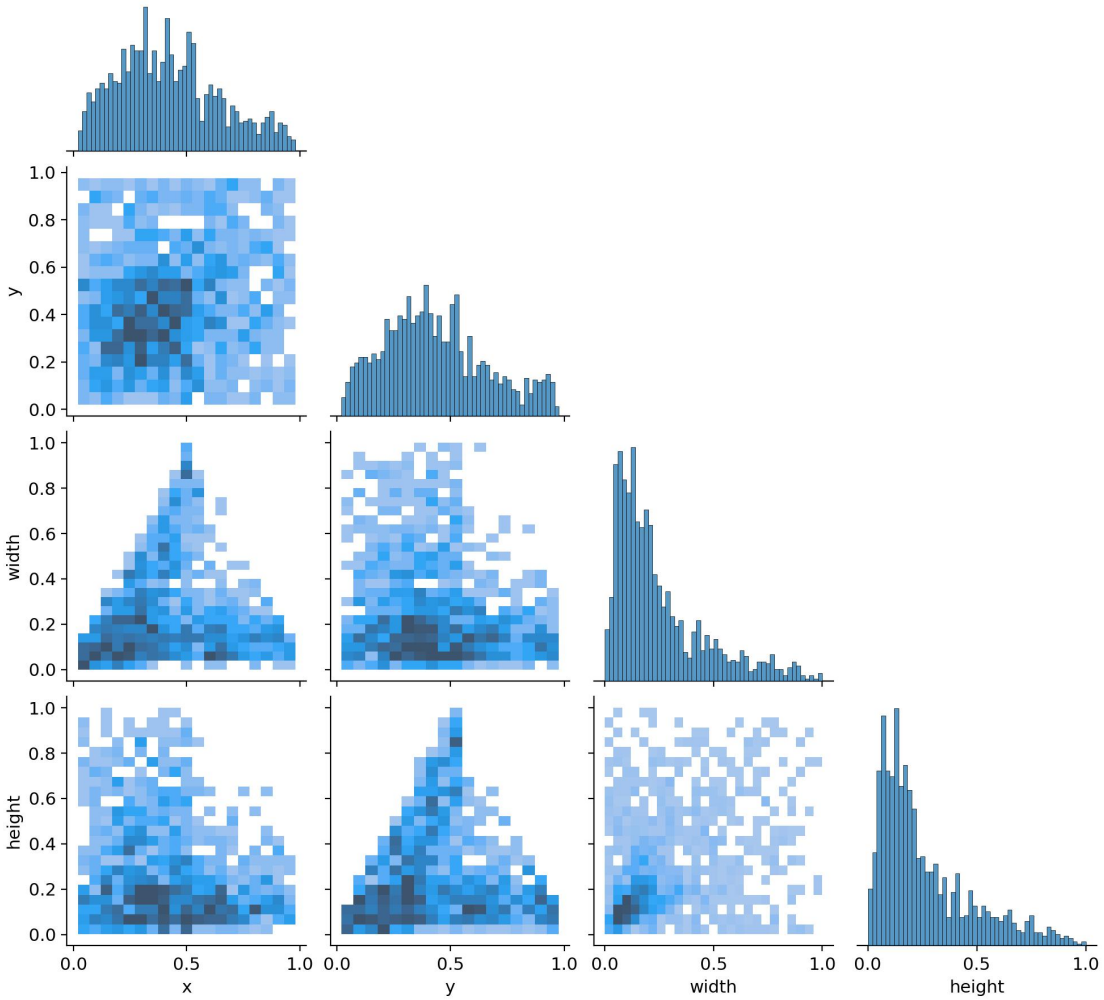

Figure 4.9: Final result of labels correlogram obtained from YOLOv5

between precision and recall at various confidence thresholds. It demonstrates the model's ability to detect all relevant objects (recall) while maintaining a high precision. The PR-curve provides insights into the model's overall performance and can help determine the optimal threshold for a specific application. The R-curve, or Recall curve, highlights the model's ability to correctly identify true positive detections across different levels of recall. It shows the relationship between the recall rate and the number of detections made by the model. A higher recall value indicates a higher proportion of true positive detections being captured.

The P-curve, PR-curve, and R-curve provide a comprehensive view of the object detection model's performance. These visualizations enable us to evaluate the precision, recall, and trade-offs associated with different

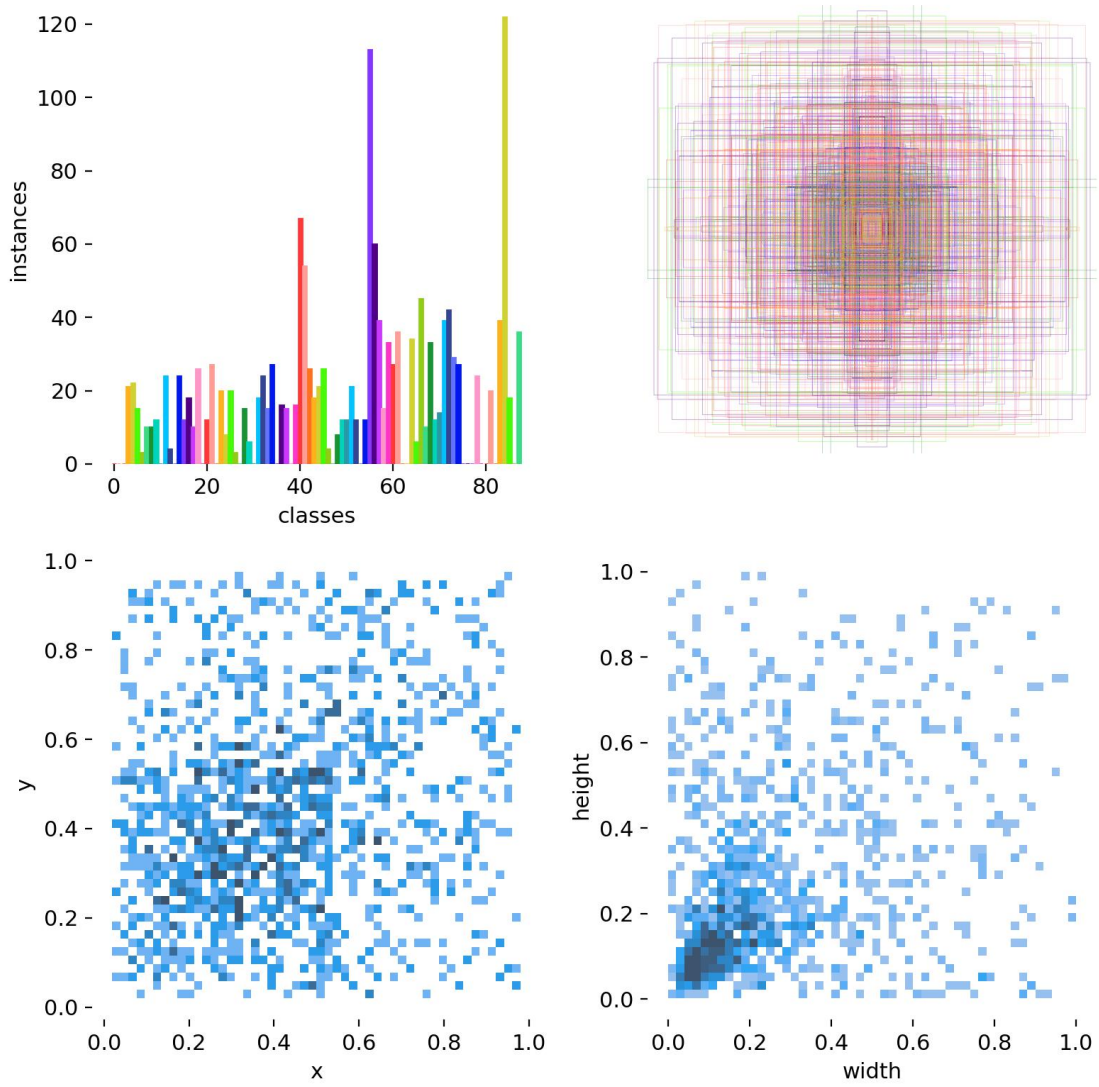

Figure 4.10: Graphical results obtained from YOLOv5

confidence thresholds. The analysis of these curves helps in making informed decisions about the model's performance and its applicability to real-world scenarios.

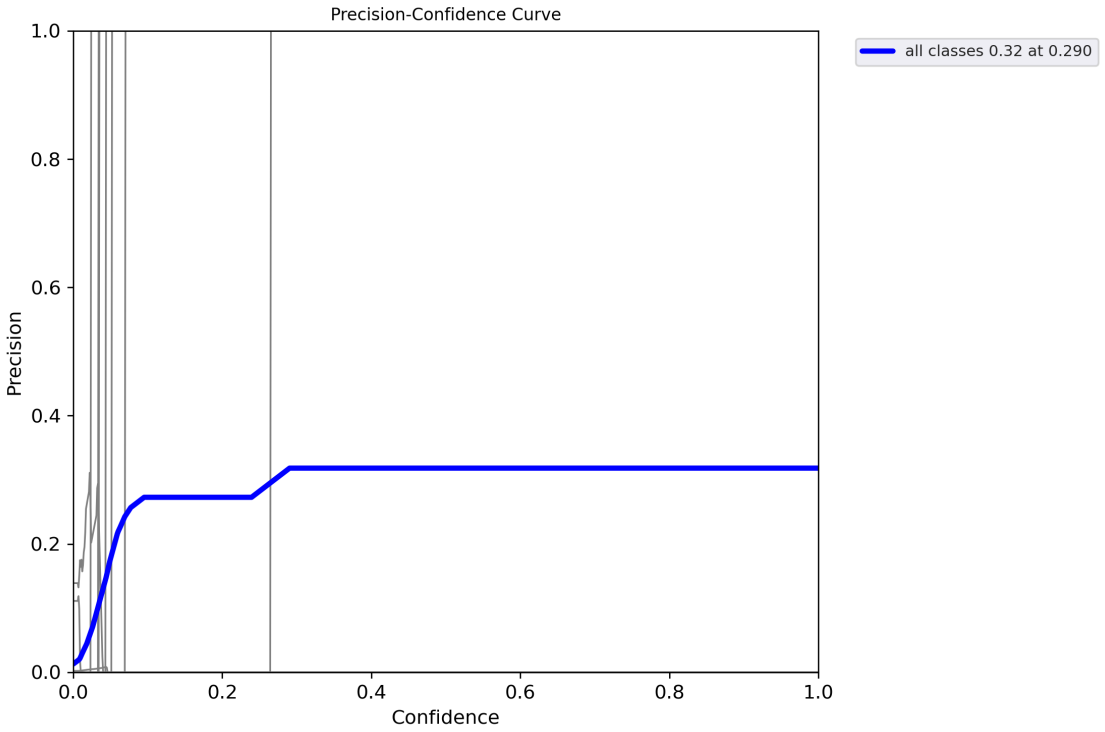

Figure 4.11: Final result of P curve obtained from YOLOv5

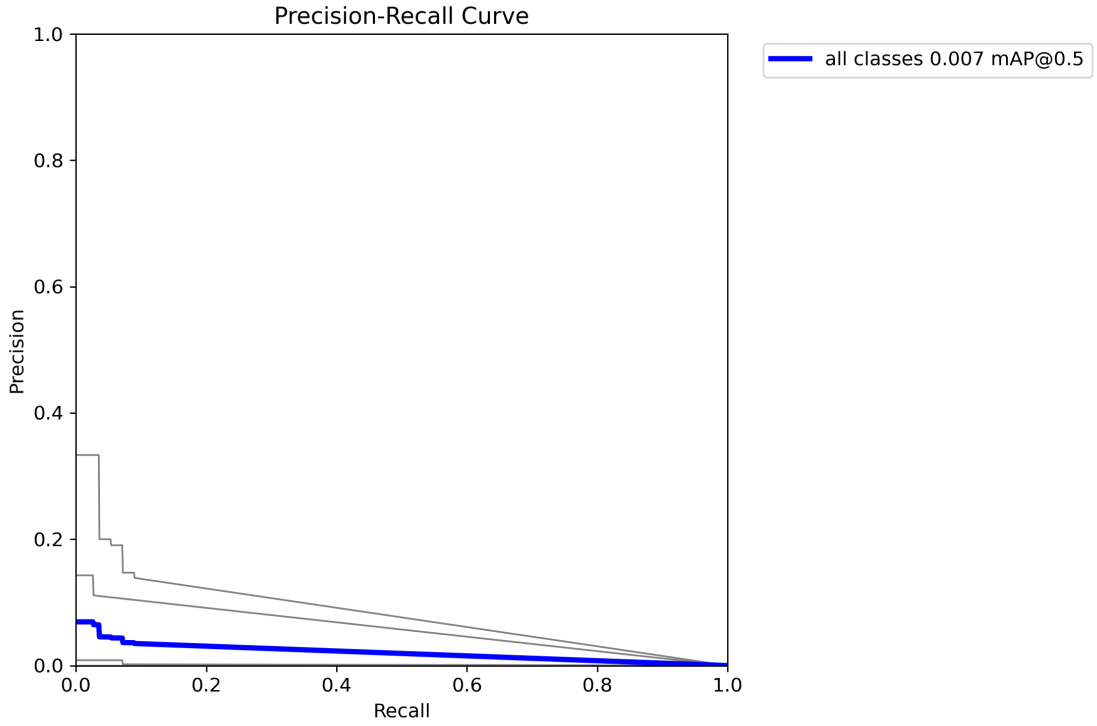

Figure 4.12: Final result of PR curve obtained from YOLOv5

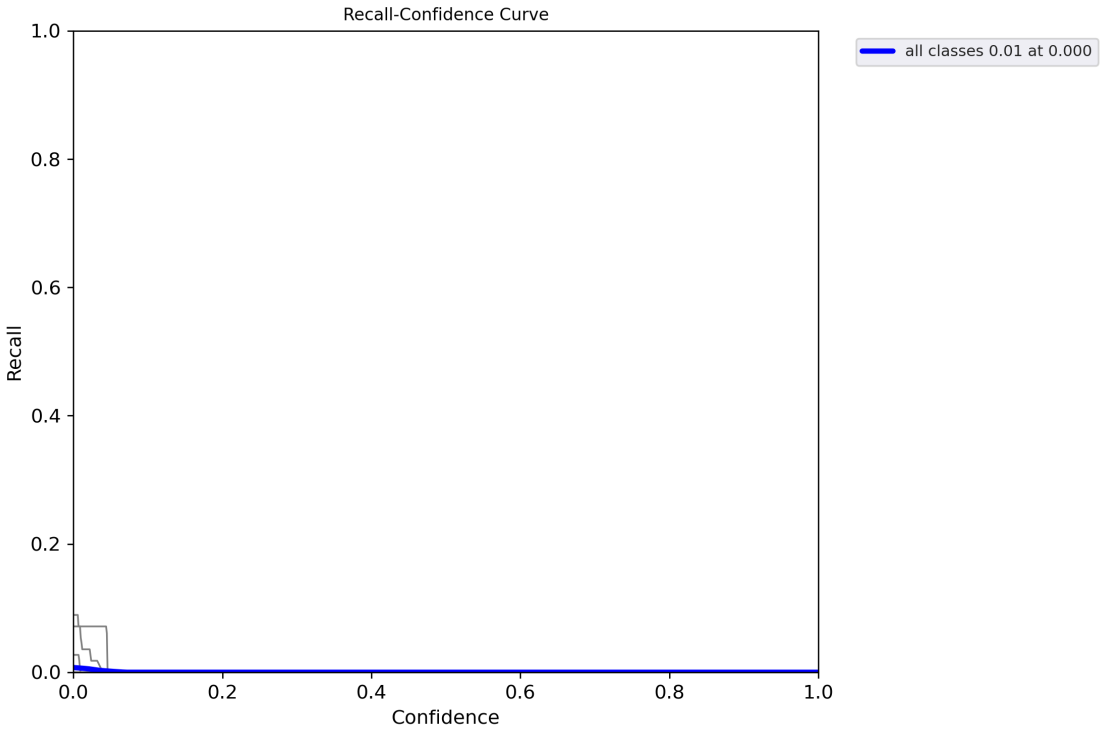

Figure 4.13: Final result of R curve obtained from YOLOv5

Ideally, we want both precision and recall to be high, indicating accurate and comprehensive detection of positive instances. However, there is often a trade-off between the two. Adjusting the threshold can increase precision at the expense of recall, or vice versa. The optimal threshold is typically chosen based on the specific requirements of the application and the desired balance between precision and recall.

Analyzing the precision/recall graph helps us assess the model's performance at different threshold settings and choose the threshold that best suits our needs. This information is crucial for evaluating the effectiveness of the model and making informed decisions regarding its deployment and performance optimization.

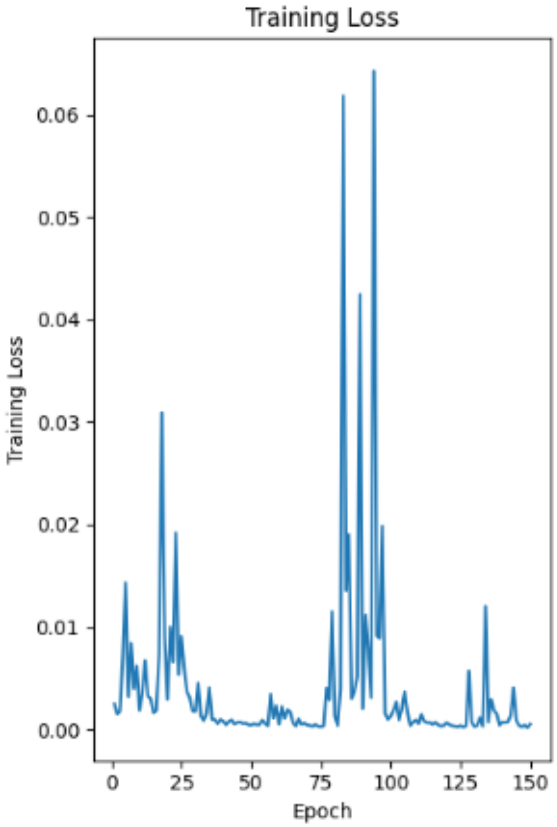

Figure 4.14: Train loss vs epoch graph of training data on BERT-base

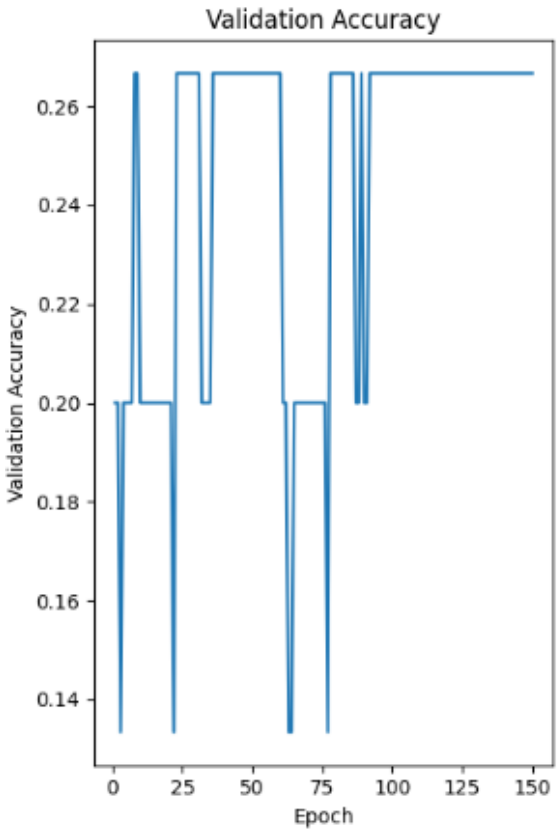

Figure 4.15: Validation accuracy vs epoch graph of validation data on BERT-base

The figure [4.14] and [4.15]shows the train loss and validation accuracy of a BERT-based model over 150 epochs respectively. The train loss first decreases and then we experienced a sharp increase and then again decreases over the epochs, indicating that the model is learning from the training data. The validation accuracy also increases over the epochs, indicating that the model is generalizing to the validation data. However, the validation accuracy plateaus at certain points indicating that the model is overfitting at certain portions due to lack of availability of training datasets.

Chapter 5

Conclusion and Future scope

5.1 Conclusion

This work focused on the development of a comprehensive system for visual question answering (VQA) by integrating YOLOv5, a state-of-the-art object detection model, with BERT, a powerful NLP model. The goal was to create a robust and accurate system capable of answering questions based on the visual content of images.

The study began by manually creating a dataset by extracting images from the book "Basic and Advanced Laboratory Techniques in Histopathology and Cytology" by Pranab Dey. The images were annotated using the Roboflow workspace platform and assigned class labels for object detection. Data augmentation techniques were applied to address the limited availability of the dataset and enhance model generalization.

Simultaneously, a novel dataset was created for question-answer pairs to link and match the features detected by YOLOv5 with their corresponding details. The YOLOv5 architecture was integrated with BERT to generate meaningful answers based on the detected objects in the images. This fusion of computer vision and natural language processing techniques aimed to provide accurate and contextually relevant responses to user queries.

The evaluation of the system was performed using standard evaluation metrics for object detection, such as mean average precision (mAP), and for question answering, metrics like accuracy and BLEU score. These metrics provided insights into the performance and effectiveness of the integrated system.

The results demonstrated the effectiveness of the proposed approach in accurately detecting objects in images and generating relevant answers to user questions. The fusion of YOLOv5 and BERT leveraged the strengths of both computer vision and natural language processing, enabling a com-

prehensive understanding of the visual content and generating meaningful responses.

Overall, this work showcased the potential of integrating YOLOv5 and BERT for building a robust visual question-answering system. It highlighted the importance of leveraging both visual and textual information to enhance the system's performance and provide users with accurate and informative answers to their queries. Further research and optimization of the models and techniques can lead to advancements in the field of VQA and improve the system's performance even further.

Future Scope

The present work lays a strong foundation for the integration of YOLOv5 and BERT in the context of visual question answering. However, there are several avenues for future research and improvements that can be explored:

1. Dataset Expansion: Although efforts were made to manually create a dataset, further expansion of the dataset can enhance the model's performance. Collecting more diverse and representative images, as well as generating a larger set of question-answer pairs, can help improve the system's understanding and ability to generalize.

2. Fine-tuning and Hyperparameter Optimization: Fine-tuning the YOLOv5 and BERT models with different hyperparameters can lead to improved performance. Exploring different learning rates, batch sizes, and regularization techniques can help fine-tune the models for better accuracy and faster convergence.

3. Multimodal Fusion: Currently, the integration of YOLOv5 and BERT focuses on sequential processing, where object detection precedes question answering. Exploring multimodal fusion techniques that enable simultaneous processing and fusion of visual and textual information can potentially improve the system's efficiency and accuracy.

4. Contextual Understanding: Enhancing the system's ability to understand and reason with contextual information can significantly improve its performance. Incorporating contextual understanding techniques, such as attention mechanisms or memory networks, can enable the model to capture dependencies and relationships between objects and generate more contextually relevant answers.

5. Transfer Learning and Pretraining: Leveraging transfer learning and pre-trained models can be beneficial for improving the performance of both YOLOv5 and BERT. Fine-tuning on larger-scale pre-trained models

or domain-specific datasets can help the models capture more intricate patterns and improve their generalization capabilities.

6. User Interaction and Feedback: Integrating user interaction and feedback mechanisms can further enhance the system's performance. Incorporating techniques that allow users to provide feedback on the generated answers or correct any inaccuracies can help refine the model's responses over time.

7. Deployment and Real-World Applications: Moving beyond the experimental setup, deploying the integrated system in real-world scenarios can provide valuable insights into its practicality and usability. Evaluating the system's performance on diverse datasets and testing it in real-world environments will contribute to its practical applicability and identify potential challenges.

Overall, the future scope for this work lies in refining the integrated YOLOv5 and BERT model, expanding the dataset, incorporating multimodal fusion techniques, improving contextual understanding, exploring transfer learning and pretraining, integrating user interaction, and deploying the system in real-world applications. These advancements will contribute to the development of more accurate and robust visual question-answering systems with practical applications in various domains.

Bibliography

[Abacha et al., 2019] Abacha, A. B., Hasan, S. A., Datla, V. V., Liu, J., Demner-Fushman, D., and Müller, H. (2019). Vqa-med: Overview of the medical visual question answering task at imageclef 2019. CLEF (working notes), 2(6).

[Al-Sadi et al., 2021] Al-Sadi, A., Al-Ayyoub, M., Jararweh, Y., and Costen, F. (2021). Visual question answering in the medical domain based on deep learning approaches: A comprehensive study. Pattern Recognition Letters, 150:57–75.

[Atrey et al., 2010] Atrey, P. K., Hossain, M. A., El Saddik, A., and Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: a survey. Multimedia systems, 16:345–379.

[Do et al., 2021] Do, T., Nguyen, B. X., Tjiputra, E., Tran, M., Tran, Q. D., and Nguyen, A. (2021). Multiple meta-model quantifying for medical visual question answering. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24, pages 64–74. Springer.

[Gaafar et al., 2022] Gaafar, A. S., Dahr, J. M., and Hamoud, A. K. (2022). Comparative analysis of performance of deep learning classification approach based on lstm-rnn for textual and image datasets. Informatica, 46(5).

[He et al., 2020] He, X., Zhang, Y., Mou, L., Xing, E., and Xie, P. (2020). Pathvqa: 30000+ questions for medical visual question answering. arXiv preprint arXiv:2003.10286.

[Jiang et al., 2022] Jiang, P., Ergu, D., Liu, F., Cai, Y., and Ma, B. (2022). A review of yolo algorithm developments. Procedia Computer Science, 199:1066–1073.

[Jocher et al., 2020] Jocher, G., Stoken, A., Borovec, J., Changyu, L., Hogan, A., Diaconu, L., Poznanski, J., Yu, L., Rai, P., Ferriday, R., et al. (2020). ultralytics/yolov5: v3. 0. Zenodo.

[Lin et al., 2022] Lin, Q., Ye, G., Wang, J., and Liu, H. (2022). Roboflow: a data-centric workflow management system for developing ai-enhanced robots. In Conference on Robot Learning, pages 1789–1794. PMLR.

[Lin et al., 2021] Lin, Z., Zhang, D., Tac, Q., Shi, D., Haffari, G., Wu, Q., He, M., and Ge, Z. (2021). Medical visual question answering: A survey. arXiv preprint arXiv:2111.10056.

[Lin et al., 2023] Lin, Z., Zhang, D., Tao, Q., Shi, D., Haffari, G., Wu, Q., He, M., and Ge, Z. (2023). Medical visual question answering: A survey. Artificial Intelligence in Medicine, page 102611.

[Liu et al., 2019] Liu, A., Huang, Z., Lu, H., Wang, X., and Yuan, C. (2019). Bb-kbqa: Bert-based knowledge base question answering. In Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18, pages 81–92. Springer.

[Liu et al., 2022] Liu, H., Sun, F., Gu, J., and Deng, L. (2022). Sf-yolov5: A lightweight small object detection algorithm based on improved feature fusion mode. Sensors, 22(15):5817.

[Lohith et al., 2023] Lohith, M., Bardhan, S., and Bandyopadhyay, O. (2023). 10 cervical pap smear screening and cancer detection using deep neural network. Current Applications of Deep Learning in Cancer Diagnostics, page 125.

[Omar et al., 2023] Omar, R., Mangukiya, O., Kalnis, P., and Mansour, E. (2023). Chatgpt versus traditional question answering for knowledge graphs: Current status and future directions towards knowledge graph chatbots. arXiv preprint arXiv:2302.06466.

[Rothman and Gulli, 2022] Rothman, D. and Gulli, A. (2022). Transformers for Natural Language Processing: Build, train, and fine-tune deep neural network architectures for NLP with Python, PyTorch, TensorFlow, BERT, and GPT-3. Packt Publishing Ltd.

[Sarvamangala and Kulkarni, 2022] Sarvamangala, D. and Kulkarni, R. V. (2022). Convolutional neural networks in medical image understanding: a survey. Evolutionary intelligence, 15(1):1–22.

[Verma and Ramachandran, 2020] Verma, H. and Ramachandran, S. (2020). Harendrakv at vqa-med 2020: Sequential vqa with attention for medical visual question answering. In CLEF (Working Notes).