Implementing Object Detection into Medical Report Generation to Improve Accuracy

Lachlan Bassi
Bachelor of Advanced Computer Science (Hons)
University of Western Australia
Course: CITS4010

I. Abstract

This study aims to enhance Medical Report Generation (MRG) for Fundus Fluorescein Angiography (FFA) images and presents a methodology with broader applicability. The central hypothesis posits that integrating noise-free, lesion-specific FFA images into MRG model training will significantly improve accuracy on unseen data.

The research's significance extends to patient care, potentially expediting diagnoses and treatments for eye conditions, reducing radiologists' workload, and ensuring timely care. Moreover, this study presents a proposed model that surpasses the baseline model from the FFA-IR paper [7].

The scope of this research focuses solely on FFA image MRG without exploring other image types or dataset labelling processes.

The research methodology consists of three parts. Part 1 employs text mining to classify lesions based on medical reports, complementing object detection. Part 2 trains an object detector to identify prevalent lesions. Part 3 involves image cropping and MRG model evaluation, demonstrating enhanced performance and validating the proposed methodology.

The research outcomes are as follows. Firstly, the formulation of a model adept at categorising medical reports into one of 46 lesion types, boasting an average recall rate of 75.43% across these categories. Secondly, an inaugural object detector, grounded in the FFA-IR dataset, was developed. It successfully identifies the four predominant lesions documented in the labelled lesion data provided by [7] which are: Macular pucker (or Epiretinal membrane), Choroidal neovascularization, High myopia choroidal neovascularization, and Lacquer crack pathological myopia macular lesion. These lesions showcased precision metrics ranging between 0.65 and 0.8 at their optimal confidence intervals. Conclusively, when the MRG model was supplied with lesion-focused image sections via the object detector, a marked uptick was observed across all evaluative parameters in the testing set-most prominently, a 9.9% surge in CIDEr accuracy compared to the baseline.

This research encourages the exploration of object detection within MRG across various datasets, not limited to FFA images, with the aim of discerning whether a consistent enhancement in accuracy can be achieved universally.

II. Acknowledgments

Thank you to Cara MacNish the Honour's Supervisor for 2023 and thank you to my supervisors Naeha Sharif and Ghulam Mubashar Hassan for your continued support and guidance throughout this project.

The code used for this project can be found at: GitHub.

III. Introduction

A. The Problem

Wait times for diagnostic imaging not only jeopardise medical outcomes, but also impose emotional and financial burdens on patients. As the demand for imaging grows, addressing this issue becomes crucial for ensuring both the quality and efficiency of healthcare [1]. Medical Report Generation (MRG), is the process of automatically generating medical reports using computer vision and deep learning. The goal is to design a model that can generate medical reports accurately taking a large workload off of radiologists, this will allow them to spend their time attending to more patients reducing the risk of patients missing out on urgent treatment. In Canada for example, a country that lacks behind other developed nations in terms of Magnetic Resonance Imaging (MRI) equipment per capita, people wait 3.7 weeks on average for CT scans and 11.1 weeks for MRI scans [1]. Long wait times have been found to cause cancer tumours to increase in size and in the same paper they mention that 62

The focus of this paper is on improving the accuracy of medical report generation of Funds Fluorescence Angiogram (FFA) images, which are images of the eyeball taken after a yellow dye has been injected into them to identify the health of the retina and blood vessels [2]. According to the International Diabetes Foundation, roughly a third of the 400 million people in the world with diabetes are believed to have diabetic Retinopathy [3]. This condition causes damage to the light sensitive tissue in the retina which can lead to vision problems and eventually blindness if untreated. Similarly, another common eye condition is age related macular degeneration (AMD) which is a leading cause of vision loss. In 2022 it is estimated that across the world there are 200 million people who have some form of AMD and this number is projected to grow to almost 300 million people by 2040 [4].

Eye conditions effect hundreds of millions across the world so improving MRG for FFA images can allow people to receive the care they need quicker. Currently MRG faces big issues in terms of accuracy [5] [6]. This study seeks to enhance a foundational MRG model introduced by the authors in the article titled ”FFA-IR: Towards an explainable and reliable medical report generation benchmark” [7]. This article was showcased at NeurIPS 2021, a leading international conference on machine learning, artificial intelligence, and computational neuroscience. The paper introduces a new dataset of FFA images specifically tailored for MRG, it distinguishes itself from existing MRG datasets due to its larger image count and longer average word length within the medical reports. In the same paper they also present a baseline model for MRG using the FFA-IR dataset and the results show serious limitations in their accuracy.

To improve the baseline model’s accuracy, this paper incorporates object detection to crop and reduce noise in images before feeding them to the MRG model for report generation. Object detection is the process of detecting and labelling an area of interest in an image, in the context of this study it is the detection and labelling of lesions in FFA images. Remarkably, the utilisation of object detection for improving MRG accuracy has been an under explored avenue, despite its successful implementation in related research domains, such as image captioning and content description [8] and [9]. In both papers the findings demonstrated a significant improvement in accuracy and this was the motivation behind this research.

B. Problem Statement

This study is the first of its kind at utilising object detection to attempt to improve Medical Report Generation accuracy in the context of FFA images. Reasons for this include that the FFA-IR paper [7] was only released in 2022, before it there wasnt a large dataset of labelled images of lesions in FFA images widely available online. Labelled images are required to train an object detector as they are ground truth values for what a specific lesion looks like. Object detectors pre-trained on general images are not suitable for this use due to the unique intricacies of medical images. By integrating object detection into medical report generation the hope is to find an improvement in accuracy just like the papers [8] and [9].

C. Purpose of the Study

The objective is to establish a more precise MRG baseline for FFA images and potentially present a methodology that can be followed for improving MRG models for other medical images.

D. Hypothesis

Hypothesis: Integrating lesion-specific images from FFA, devoid of extraneous noise, into training the MRG model will enhance its accuracy on unseen data.

E. 5. Significance of the Study

The significance of this study extends beyond the academic realm, touching the very core of patient care. By enhancing the accuracy of Medical Report Generation (MRG) for FFA images, the potential to expedite diagnosis and treatment for millions suffering from eye conditions becomes a tangible reality. This research could pave the way for more efficient uses of radiologists time, potentially reducing wait times and ensuring that patients receive timely care. Furthermore, the implications of this study could ripple across the medical imaging field, setting a precedent for integrating object detection in various other imaging modalities.

F. Scope of the Study

This research will focus primarily on the integration of object detection in MRG for FFA images. While the findings may have implications for other areas of MRG, the study will not delve into the application of this methodology in other imaging types. Additionally, while the study acknowledges the importance of labeled images in training object detectors, the process of labelling or the intricacies of creating such datasets will not be covered.

G. Brief Methodology

The research methodology encompasses three pivotal facets. Initially, the FFA-IR dataset's labeled bounding box data will be used to train an object detector. Next, this detector will identify lesions in the remaining FFA images, cropping them to exclusively showcase the lesion and removing peripheral noise. The final step involves comparing the accuracy of the MRG model to the baseline and discern if there are any improvements.

H. Dissertation Structure

Chapter 4 delves into a detailed literature review, exploring the evolution of MRG, the significance of FFA images, and the role of object detection in image analysis. Chapter 5 outlines the research methodology in detail, discussing the rationale behind chosen methods and potential challenges. Chapter 6 presents the research findings, offering a comparative analysis of MRG accuracy with and without the integration of object detection. Chapter 7 discusses the implications of the findings, drawing connections to existing literature and suggesting areas for future research. Chapter 8 concludes the dissertation, summarising the key takeaways and their significance in the broader context of medical imaging and patient care.

IV. Literature Review

A. Structure of Literature Review

This literature review is structured into specific segments, encompassing an exploration of the FFA-IR dataset, an analysis of Medical Report Generation, a review of evaluation metrics pertinent to medical report generation, an examination of object detection in the context of FFA images, a discussion on evaluating an object detector's efficacy, and culminating with an insight into categorisation via text mininga method adopted in the methodology as a response to certain limitations in the dataset.

B. The FFA-IR dataset

The FFA-IR data set [7] is a collection of FFA images from patients admitted to the Zhongshan Ophthalmic Center of Sun Yat-sen University in Guangzhou, China. This data set contains 10,790 medical reports along with 1,048,584 FFA images. The data set also contains explainable annotations of 46 categories of lesions with a total of 12,166 regions. The paper conducted a comparative study between FFA-IR and 9 other prominent and readily accessible MRG datasets. These were Open-IU, DEN, COV-CTR, MIMIC-CXR, PadChest, CX-CHR, STARE, DIARETDB1, and MESSIDOR. FFA-IR distinguished itself from the rest due to it having the largest number of medical images (1,048,584) and the longest average length of reports (91.2 words) see figure 1.0.1 (appendix). Moreover, FFA-IR offers interpretable annotations by labelling 46 types of lesions across 12,166 regions accompanied by corresponding FFA images and reports, which are essential for identifying diseases and report generation. The data set includes bounding box co-ordinates for lesions and contains a medical report for each image, which are features that can be used for training a detection model.

C. Medical Report Generation

1) Initial Approaches: The roots of MRG can be traced back to several early methodologies. In 2012, Varges et al. [11] used natural language generation to produce doctors letters. The approach required the doctor to enter the observations in text form where the system would then generate the report.

This technology was primarily applied in cardiology, where it was used to document findings related to heart conditions. The system worked by utilising two main components: Medical ontology: this a structured set of medical terms and relationships that help the system understand and organise the information input by the doctor. Corpus-informed syntactic frame approach: this is a method of language processing that uses a large body of text (corpus) to inform the structure (syntax) of the generated report. Essentially, the system learns how to structure sentences and organise information in a way that mirrors existing medical texts. The reports were generated to conform to the HL7 Clinical Document Architecture Standard, a framework for structuring clinical documents, and was then evaluated by medical staff.

In 2015 [12] proposed a method using convolutional neural networks (CNNs), a type of deep learning architecture specifically designed for processing grid-like data such as images, to predict semantic descriptions from medical images. They applied their method to retinal images obtained from OCT scans which are used to diagnose and monitor eye diseases such as age-related macular degeneration and diabetic retinopathy. By demonstrating the potential of CNNs to generate meaningful and accurate semantic descriptions of medical images the authors paved the way for further research and applications of deep learning in medical image analysis.

Following this, in 2016 a multi-task loss CNN architecture for medical image captioning was proposed [13]. This approach combines two related objectives: image classification and image captioning. The model is trained to perform both tasks simultaneously, which allows the model to leverage shared features between the two tasks improving its overall performance. This multi-task learning framework showcased a promising approach for handling image classification and captioning tasks in the medical imaging domain.

Despite the efforts of these prior methods, none achieved substantial success. Instead, it was an approach proposed by Cho et al. in 2014 [14] that proposed an encoder-decoder architecture that would later go on to become a standard for Medical Report Generation (MRG).

2) Encoder-Decoder Architecture: In the research report, Cho et al. [14], proposed an encoder-decoder architecture for the application of statistical machine translation, the process of converting one language such as English to another such as French. The architecture consisted of a recurrent neural network (RNN) encoder and a RNN decoder. Since then this architecture has been adapted to be used in many other domains including MRG [5].

The paper [15] is one of the earliest works that applies an encoder-decoder architecture for MRG. It uses a CNN-RNN architecture instead of an RNN to RNN structure because CNNs are better at interpreting images as they are designed for grid-like data such as images where spatial structure and local patterns are important [5]. RNNs [16] on the other hand are specifically designed to process sequences of data such as time-series, natural language, and other data that has a temporal or sequential structure.

While CNNs are used for encoding the RNN network is still generally used for decoding. In [15] the RNN used is called a Long Short-Term Memory (LSTM) [16], which is according to the survey paper, Diagnostic captioning: a survey [5], is a very common choice of RNN.

LSTM [16] was proposed in 1997 and it is a type of RNN designed to address the vanishing gradient problem commonly encountered in training RNNs. The vanishing gradient problem occurs when gradients of the loss function with respect to the networks weights become very small during back propagation. This issue makes it difficult for RNNs to learn and capture long-range dependencies in the input sequences, limiting their effectiveness in handling tasks that involve longer sequences. LSTM overcomes this limitation by incorporating specialised memory cells and gating mechanisms that allow the network to store and access information over long periods. The result is LSTMs are able to learn long-range dependencies and perform better on a wider range of sequence-to-sequence tasks.

Although LSTMs and other RNNs continue to be a staple in the design of encoder-decoder architectures for MRG, the rise of Transformer models have prompted a significant shift in contemporary research. Increasingly, these Transformer models are being utilised for both encoding and decoding tasks, setting a new direction in the field of MRG and beyond.

3) Transformer encoder - decoder: Transformers [17] are a type of neural network architecture and are built upon the concept of self-attention mechanisms. This feature allows the model to assign varying degrees of importance to different input tokens based on their relative significance within a given context. The mechanism helps the model capture complex dependencies and relationships within the input sequence. The transformer architecture [17] uses an encoder-decoder framework, each composed of multiple layers of self-attention and feed-forward neural networks.

The encoder processes [17] the input sequence of tokens and creates a contextualised representation for each token. This means that each token is represented in a way that reflects its meaning in the context of the rest of the sequence. For example, the word eagle has different meanings in golf and for birds and so using the context of the sequence the transformer determines the meaning.

Following the encoding process the decoder [17] then proceeds to construct the output sequence one token at a time. As it goes about this, the decoder draws on two key sources of information. The first is the encoders output, the decoder uses the contextual representations of the tokens generated by the encoder. The representations, as explained earlier, reflect the meaning of each token in the context of the entire input sequence. So, for example, if the word eagle in the input sequence was made clear to be in the context of birds the decoder would use that information as it generates information. The second is the self-attention mechanism, which allows the decoder to consider the relationships between the tokens it has already generated in the output sequence as it decides what the next token should be. This helps ensure the tokens in the output sequence make sense in relation to each other, much like how the words in a sentence need to make sense together.

Transformers have been shown to outperform CNN-RNN models when it comes to MRG in certain medical benchmarks [18]. The advantage they have is where RNNs process input sequences one item at a time transformers process the entire input sequence of data at once using the mechanism of self-attention. This allows transformers to capture long-range dependencies more effectively then RNNs [18].

4) Retrieval Approaches for MRG: Another approach is known as the Hybrid Retrieval-Generation Reinforced Model (HRGR) [19]. This model leverages both a retrieval-based module and a generation-based module. The retrieval-based modules role is to locate relevant medical reports from a large corpus of pre-existing reports. Subsequently, the generation-based module takes into account the visual features extracted from the medical image to generate the final report.

The survey paper, Diagnostic Captioning: a survey [5], showed how in the act of diagnostic captioning, the process of generating a descriptive diagnostic report or caption for a given medical image, that an elaborate retrieval based system [20] outperformed state of the art encoder-decoders.

5) Summary: In the context of Medical Report Generation (MRG), the strengths and potential applications of each methodology are nuanced. Transformer models, with their ability to process complete sequences simultaneously and effectively capture intricate relationships within the data, are currently seen as the most promising route for progress in MRG [6]. Hybrid models, such as the HRGR, are compelling contenders as they adeptly synthesise the best of different methodologies, and in the future could potentially usher in a new era of integrated solutions. The field of MRG is dynamic and fast-paced, with constant exploration and testing of new models and architectural concepts. Consequently, the ”best” method remains fluid, expected to evolve in tandem with advancements in the domain.

D. Medical Report Generation Evaluation

1) Overview: Evaluating the performance of an MRG model can be a challenge without access to a medical professional to critique each report individually. Fortunately there exist multiple Natural Language Generation (NLG) evaluation tools developed for this purpose. BLEU [21], CIDEr [22], METEOR [23], and ROUGE [24] aim to calculate the similarity between source and target sentences based on the occurrences of N-gram or word matching.

2) NLG Evaluation Systems: **BLEU:** Bilingual Evaluation Understudy (BLEU) [21] is a metric designed to assess the similarity between a candidate sentence (often a machine-generated translation) and one or more reference sentences. It gauges this similarity based on the overlap of n-grams contiguous sequences of n items from a given sample of text between the candidate and the references. BLEU can be computed at various n-gram levels:

BLEU-1: Focuses on the overlap of unigrams (single words). BLEU-2: Considers both unigrams and bigrams (two-word sequences). BLEU-3: Incorporates unigrams, bigrams, and trigrams (three-word sequences). BLEU-4: Extends the comparison to include unigrams, bigrams, trigrams, and four-grams (four-word sequences).

The BLEU score ranges from 0 to 1, with higher scores signifying greater similarity between the candidate and reference sentences. It's noteworthy, however, that BLEU has its limitations. One primary limitation is its strict reliance on exact matching of n-grams, which can overlook potential synonyms or alternative phrasings that might be contextually accurate. This exactness means that even semantically correct translations can receive lower scores if they don't match the reference phrasings precisely.

CIDEr: Consensus-based Image Description Evaluation (CIDEr) [22] is another metric used to evaluate image captioning models. CIDEr calculates the similarity score between the candidate and reference sentences. CIDEr takes into account the diversity of human-generated sentences and captures a wider range of possible correct captions. Its limitation, though, is the requirement for a significant number of human annotators to yield reliable results, a condition satisfied by the FFA-IR data set.

METEOR: Metric for Evaluation of Translation with Explicit Ordering (METEOR) [23] evaluates the similarity between candidate and reference sentences based on a weighted combination of unigram precision, recall, and an alignment-based penalty for incorrect word order. Its designed to be more robust for differences in word order and syntax between candidate and reference sentences. However, it demands an alignment step between candidate and reference sentences, which can be computationally expensive.

ROUGE: The Recall-Oriented Understudy for Gisting Evaluation (Rouge) [24] measures the similarity between candidate and reference sentences based on the overlap of n-grams and word sequences. Rouge is commonly used in summarisation tasks as it measures the recall relative to the reference. A limitation of Rouge is that it doesnt take into account the fluency or coherence of the candidate summary, which is important in a Medical Report.

3) Summary: Based on a survey of Transformer-based MRG models [6], the majority use both CIDEr and BLEU. CIDEr and BLEU excel in scenarios where the precision and diversity in generated sentences are important [6]. Comparatively, METEOR and ROUGE demonstrate strengths where word order and recall are crucial. Consequently, integrating ROUGE or METEOR alongside BLEU and CIDEr could be beneficial to an assessment of generated medical reports as there will be a better account for recall.

E. Detection Algorithms

1) Overview of Object Detection: Object detection [25] is a task in computer vision that involves identifying and locating objects of a certain class (like people, cars, or animals) in digital images or videos. Object Detector CNNs are one type of model used to perform this task.

In the baseline R2Gen code [10], a CNN is employed for feature extraction, with the extracted features subsequently passed to a transformer for MRG. This research endeavours to first process the images through an object detector, allowing for the images to be cropped to display only the lesion. These cropped images are then directed to the feature extractor.

Object Detector CNNs [25] dont just classify an image as containing an object or not, as in image classification tasks. Instead, it also provides a bounding box that locates where in the image the object is. This is done by a process called localisation and combines with classification in a single model to perform object detection.

2) Two Stage Approaches: When it comes to object detection there are two methods: two stage approaches and one stage approaches. A common two-stage approach is Faster R-CNN and is built off of R-CNN and Fast R-CNN [26].

3) A brief history of Faster R-CNN: R-CNN and Fast R-CNN: **R-CNN** [27] (Region-based Convolutional Neural Network) is a deep learning approach to object detection that was developed by researchers at UC Berkeley in 2014. R-CNN is capable of identifying up to 80 distinct objects within images, this was one of the first object detection models that employed a CNN to extract features from images.

The approach consists of three main modules: the first module uses the Selective Search Algorithm to generate around 2000 region proposals which are networks of bounding boxes of an input image that might have objects within them, the second module extracts feature vectors for each region proposal using a pre-trained CNN, and the third module classifies the region proposals using a pre-trained, Support Vector Machine, SVM algorithm which determines if they are either the background or one of the object classes. Since SVMs are a binary classifier there is one SVM for each object class meaning there are N + 1 SVMs trained in total, where N is the number of object classes and the plus one is for the background class, each is tested and the class with the highest score is used to classify the object.

Despite its success R-CNN has some drawbacks. Being a multi-staged model, it lacks the ability to be trained end-to-end. Other drawbacks include it requiring large storage to cache the SVM-extracted features, and its inability to run in real-time.

Fast R-CNN [28] was proposed by Ross Girshick in 2015 and it overcomes several issues with R-CNN, primarily increasing its speed and making it more computationally efficient. Fast R-CNN can be broken down into the following key components.

The first, like R-CNN, Fast R-CNN relies on an external algorithm, such as Selective Search, to generate region proposals. The next stage uses a pre-trained CNN for feature extraction, but instead of processing each region proposal independently for an image it extracts features from the entire input image at once, significantly reducing computation time. Next it uses Region of Interest (ROI) Pooling, which takes the feature maps generated by the CNN and extracts fixed-size feature vectors for each region proposal. From here, Fast R-CNN uses a fully connected layer followed by a softmax layer to classify the region proposals into object classes and a background class. This is a departure from R-CNN, which used SVMs. Finally it implements bounding box regression, working alongside the classification layer, this layer refines the coordinates of the region proposals to more accurately fit the objects.

The result of integrating the feature extraction and classification steps into a single network and sharing computation across proposals allows Fast R-CNN to significantly improve the speed and efficiency of the object detection process compared to R-CNN.

Despite the benefits, Fast R-CNN continues to rely on the Selective Search algorithm, a process that is notably time-consuming, for generating region proposals. The inability to tailor the Selective Search method to suit specific object detection tasks can potentially lead to inaccuracies in identifying all relevant objects within a dataset. Additionally, despite its improvements, Fast R-CNNs speed still doesnt operate at the speed of real-time applications.

Faster R-CNN [29] was also proposed in 2015 and is an object detection model that extends the Fast R-CNN architecture. It introduces the Region Proposal Network (RPN), which is a fully convolutional network that generates region proposals with various scales and aspect ratios. This is done by using anchor boxes, which are reference boxes of specific sizes and aspect ratios that are placed at different positions throughout an image. The advantage of anchor boxes is that it allows the model to detect objets at different scales and aspect ratios without having to use pyramids of images or pyramids of filters which speeds up computation time and simplifies the architecture.

Faster R-CNN is broken into the RPN and the Fast R-CNN which leads to a reduction in overall processing time. As a two stage model, first it employs the RPN to focus attention on potential object-containing regions within the image. These areas are then parsed to the Fast R-CNN for further processing and object detection.

The RPN in Faster R-CNN distinguishes itself from earlier approaches, such as Selective Search, by utilising a fully convolutional network to generate region proposals. The RPN operates on the image by using the same convolutional layers as the Fast R-CNN detection module. This shared architecture allows for end-to-end training and allows for customisation to the specific detection task, resulting in superior region proposals. Also by sharing the convolutional layers between the RPN and Fast R-CNN they are integrated into a single network enhancing the efficiency of the training process.

4) Training Faster R-CNN: The creators of Faster-RCNN [29] mention three ways of training both the RPN and Fast R-CNN while sharing the convolutional layers, these are: Alternating Training, Approximate Joint Training, and Non-Approximate Joint Training. Alternating is the most commonly used method and is preferred by the creators.

5) Alternating Training: The RPN training initiates the process, generating region proposals using shared convolutional layers. These layers weights are initialised based on a pre-existing model trained on ImageNet, while the remaining RPN weights are randomly initialised.

Once the RPN has been trained and region proposals generated, the weights of the RPN and the shared convolutional layers are fine-tuned. These tuned weights are then used to initialise the shared convolutional layers for the next stage - training the Fast R-CNN network. The remaining Fast R-CNN weights and the weights of the shared layers are tuned. Upon completion of the Fast R-CNN training, the tuned weights of the shared layers are used once more to train the RPN, and the cycle repeats.

6) Approximate Joint Training: The second method employed in Faster R-CNN training is referred to as approximate joint training. In this approach, the RPN and Fast R-CNN are treated as a single integrated network, rather than as separate modules. This means that region proposals are directly generated by the RPN, and without any immediate update to the weights of either the RPN or the shared layers, these proposals are fed into the Fast R-CNN for object location detection. The weights within the Faster R-CNN network are only tuned after the Fast R-CNN has generated its outputs.

This methods gradients of the weights with respect to the region proposals are disregarded since the shared layers and RPN weights are not updated immediately after the region proposals are produced. This results in reduced accuracy in comparison to the alternating training method [29]. However, it does bring about a 25-50

7) Non-Approximate Joint Training: In contrast, Non-Approximate Joint Training makes use of an ROI (Region of Interest) Warping layer. This layer allows for the calculation of the weights gradients with respect to the proposed bounding boxes, enhancing the model's capacity for precision in object localisation. This method is computationally the most expensive and only provides a small increase in accuracy.

8) One Stage Approaches: Unlike two-stage approaches that seperate object localisation and classification into distinct steps, one-stage approaches, such as SSD [30] and YOLO [31], perform these tasks simultaneously. They directly predict the object class and bounding box co-ordinates from the input image in a single pass, resulting in faster processing times but often at the expense of accuracy, especially for small or complex objects [26].

Single Shot MultiBox Detector (SSD) [30] is an object detection algorithm that simultaneously detects and classifies objects in images. It is a popular and efficient method for real-time object detection tasks, as it can identify multiple objects in a single forward pass through the neural network. SSD is built on a CNN and uses a combination of anchor boxes and feature maps to detect objects at various scales and aspect ratios. The main components can be summarised as a base network, multi-scale feature maps, default bounding boxes and prediction layers.

The foundation of SSD lies in a pre-trained CNN base network such as ResNet or VGG-16 which is used to extract features from the input image. The output is a feature map that preserves spatial information while reducing the dimensions of the image.

The next component in the model incorporates Multi-scale Feature Maps. By appending several convolutional layers to the end of the base network, SSD can detect objects at different scales and sizes.

Default Bounding Boxes, also known as Anchor Boxes, are generated by SSD for each cell of the feature map. These anchor boxes have different aspect ratios and sizes allowing the model to detect objects with varying shapes and scales.

Lastly, the model includes Prediction Layers. For each anchor box, the SSD model predicts two things: the class probabilities (object classification) and the bounding box offsets (object localisation). These predictions are made using the convolutional layers with the appropriate number of filters.

SSDs are most commonly trained using end-to-end training using back propagation and stochastic gradient descent (SGD) or another suitable optimisation algorithm.

You Only Look Once (YOLO) [31] is a real-time object detection algorithm that has gained popularity due to its speed and efficiency. Like SSD, YOLO is an end-to-end deep learning model that simultaneously predicts bounding boxes and class probabilities for objects in an image. It is based on CNNs and designed for fast processing, making it suitable for real-time object detection.

Instead of creating region proposals and then classifying them, YOLO divides the input image into a grid. Each cell in the grid is responsible for predicting a fixed number of bounding boxes. For each bounding box, the model predicts co-ordinates, dimensions, confidence scores (measures the probability that a bounding box contains an object), and the class probabilities. Then these predictions are combined into a final output. Training YOLO is the same as SSD as both use end-to-end training using back propagation and SGD.

Two stage approaches tend to be more accurate compared to one-stage approaches which tend to be faster [26], [25].

9) Object Detection in Medical Studies: In medical contexts, two-stage models are often favoured due to their emphasis on accuracy. They have been utilised across various domains, with image captioning being particularly relevant to this study. Object detection has been leveraged to support the generation of detailed captions for medical images. For instance, a study [8] produced captions for ultrasound images by integrating an object detector at the outset to minimise noise and zero in on the region of interest. This incorporation led to a 1

Similarly, another study [9] employed object detection to enhance the accuracy of image captioning, specifically on the COCO dataset, which comprises of 300,000 images capturing a range of general objects and environments. This approach led to improvements in various evaluation metrics. They opted for Faster RCNN, valuing its emphasis on accuracy.

Within the realm of integrating object detectors with Fundus images, two notable papers have made strides: Morphological Rule-Constrained Object Detection of Key Structures in Infant Fundus Image [32] and Customised Artificial Intelligence Toolbox for Detecting Diabetic Retinopathy with Confocal Truecolor Fundus Images Using Object Detection Methods [33].

10) Morphological Rule-Constrained Object Detection of Key Structures in Infant Fundus Image. By Yinsheng Zhang and colleagues 2023.: The research underlines the importance of accurately identifying the optic disc and macula in Retinopathy of Prematurity (ROP) diagnosis using object detection. Specifically, they employed the Faster R-CNN algorithm for this detection task. To enhance the precision of the Faster R-CNN, the team introduced domain-specific morphological rules tailored to the unique characteristics of the fundus. These rules encompass restrictions based on number, size, distance, angle/slope, and position, acting as refined criteria for the detection mechanism. For example, the system understands that there should be at most one optic disc and macula and that they typically lie on the same horizontal line. When tested on a dataset of 2953 infant fundus images, the incorporation of these morphological rules led to a substantial improvement in the object detection accuracy for the macula, which increased from 0.719 to 0.811, showcasing the effectiveness of integrating morphological rules with the Faster R-CNN in the detection process.

11) Customised Artificial Intelligence Toolbox for Detecting Diabetic Retinopathy with Confocal Truecolor Fundus Images Using Object Detection Methods. By Prasanna Venkatesh and colleagues 2023.: The study introduced a unique convolutional neural network method aimed at detecting diabetic retinopathy (DR) in confocal high-resolution fundus images. This AI tool not only identifies and classifies DR but also pinpoints specific clinical signs like micro-aneurysms, hard exudates, and neovascularisation using custom annotations. Utilising the You Only Look Once (YOLO) 5 algorithm, the research encompassed 8,000 fundus images, dividing them for training, validation, and testing. The results displayed a significant accuracy increase, achieving up to 91% in predicting DR's diagnosis, severity, and related clinical signs. Notably, the object detector had an overall sensitivity of 81.6% and a specificity of 100%. This is the pioneering study to train and annotate every clinical sign of DR in high-resolution fundus images.

12) Summary: Both of the previously mentioned studies employed datasets comprising thousands of images and achieved accuracy rates ranging between 80% and 95%. However, their primary focus was on detecting only a few types of lesions or areas of interest. This raises concerns about their accuracy when faced with Fundus images that harbour lesions resembling their target, but are inherently different. In this research, the aim is to detect 46 distinct lesions using an object detector on a dataset of 12,166 images. This averages to approximately 265 images per lesion type. Given this distribution, achieving the high accuracies reported by the other two studies might be challenging. Notably, one study employed a one-stage object detector, while the other used a two-stage approach. This suggests the potential benefit of exploring multiple object detectors for this task and comparing their outcomes.

F. Detection Evaluation

1) Overview: In a literature survey on the performance metrics for object detection [26], the paper identified that the most common forms of measuring the performance of object detection algorithms was by assessing the following: a confusion matrix, precision, recall, F1 score, and mean Average Precision.

2) Confusion Matrix: A confusion matrix [26] consists of the total number of true positives, true negatives, false positives and false negatives in a 2x2 grid. To define a true positive for object detection a metric called Intersection over Union (IoU) is used which indicates the overlap of the predicted bounding box co-ordinates to the ground truth box. Higher IoU indicates the predicted bounding box co-ordinates closely resemble the ground truth box co-ordinates. Setting a threshold value for the IoU is used to determine what is a true positive, for example, if the threshold value was 0.5 then if an IoU is equal to or greater than this value then the result is considered a true positive.

3) Precision and Recall: Two metrics are often calculated using the results from a confusion matrix and they are called precision and recall [26].

The precision is defined as:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (1)$$

Recall is defined as:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (2)$$

High precision indicates the model has a low rate of false positives, meaning it accurately identifies relevant objects without many false detections. Recall, also known as sensitivity, is the ability for a model to find all relevant cases in the data. High recall indicates the model has a low rate of false negatives meaning it effectively detects all relevant objects in the data set. In practice, there is often a trade-off between precision and recall and the ideal model should balance both metrics to achieve high overall performance [26].

4) Precision-Recall Curve: To determine an appropriate balance between precision and recall a precision-recall curve [26] can be used. This is created by plotting the precision (y-axis) against the recall (x-axis) for different confidence threshold values. Where a confidence threshold value determines the confidence score that should be used to classify a certain object. For example, if confidence threshold is set to 0.6 then only the predicted bounding boxes with confidence scores equal to or greater than 0.6 will be considered as valid detections.

5) F1 Score: To determine the quality of the precision and recall the F1 score [26] can be calculated. F1 score is a metric that combines precision and recall into one value for assessing the performance of a classification model. The F1 score is the harmonic mean of the precision and recall and gives equal importance to both. The F1 score is defined as:

$$\text{F1 score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \tag{3}$$

F1 score ranges from 0 to 1 with 1 indicating a perfect precision and recall and 0 indicating worst possible performance.

6) mean Average Precision: Using the precision and recall values the mean Average Precision (mAP) can be calculated which is the main form of evaluating the accuracy of a model. It is calculated by taking the Average Precision (AP) for each object class and then calculates the mean of these values across all classes. AP is calculated by computing the area under the precision-recall curve for a specific object class.

$$\text{mAP} = \frac{1}{n} \sum_{i=1}^{n} \text{AP}_i \tag{4}$$

In this equation:

- mAP is the mean Average Precision
- n is the number of queries
- AP$_i$ is the Average Precision for the i-th query

7) Method of Implementation: To implement these methods into the current pipeline provided by the FFA-IR article [7], a library needs to be imported. The model is built on pyTorch and within pyTorch a library exists called Detectron2 which is commonly used to calculate precision, recall, and mAP for detection models built on pyTorch. An example of which is in [34] where Detectron2 is used to evaluate the performance of object detection for autonomous vehicles.

G. Text Mining in Categorisation of Medical Conditions

Text mining, which is the extraction of meaningful information from vast swathes of textual data using natural language processing, has found applications in categorising medical reports based on the presence or absence of specific conditions. The following are two recent research papers that have effectively employed text mining methods to classify medical conditions with very high accuracy.

1) Automated Classification of Free-Text Radiology Reports: Using Different Feature Extraction Methods to Identify Fractures of the Distal Fibula by Dewald CL and colleagues [35]: This research focused on classifying unstructured radiograph reports to identify fractures of the distal fibula using various text representation methods. Employing a dataset of 3268 radiograph reports, the study used methods like Bag-of-Words (BOW), TF-IDF, and doc2vec, among others, to convert free-text reports into machine-readable vectors. These vectors were then used to train models like neural networks, SVM, and logistic regression. The BOW model emerged as the most effective feature extraction method, with an accuracy of 0.97 and an AUC of 0.98.

2) Automatic Classification of Medical Reports, the CIREA project by Mustafa A and colleagues [36]: The CIREA project sought to automate the classification of diseases from textual medical reports by proposing an automatic ICD coding approach. This was driven by the cumbersome process of selecting from 52,000 pathology codes in the ICD-10, a task often faced by practitioners. The study introduced novel algorithms like the EDA desuffixer and the CLO3 classification algorithm. The researchers aimed to further enhance the project by exploring linguistic treatments and evaluating the contribution of different sections of hospitalisation reports to diagnosis definition.

3) Comparison: Both papers underscore the utility of text mining techniques in medical categorisation. The first paper, focusing on radiology reports, utilised a broader range of text representation methods and machine learning models, with the BOW method emerging as the top performer. On the other hand, the CIREA project, while also seeking to classify medical conditions, centred its efforts on automating ICD coding. It introduced unique algorithms and highlighted the potential of exploring linguistic treatments and the contributions of various report sections to diagnosis.

4) Conclusion: Text mining offers a versatile tool for categorising medical conditions. As evidenced by the two studies, diverse methodologies can be employed depending on the specific application and the nature of the textual data at hand. Whether one is discerning fractures from radiograph reports or navigating the labyrinthine world of ICD codes, text mining techniques can streamline processes, enhance accuracy, and drive more informed medical decisions.

V. Methodology

1) Introduction: The primary aim of this research is to enhance the medical report generation process by isolating lesions in FFA images. The FFA-IR paper uses a MRG model called R2Gen [10] and this serves as the benchmark to compare the performance of their baseline model with the proposed model in this paper.

2) Dataset Limitations: The FFA-IR data set, while promising 12,166 bounding boxes, provided only 6,254. Given the vastness of the data set, modifying 6,254 images to represent lesions across 46 categories would mean only 0.89% of the total number of images in the training set would be altered. Therefore, a larger sample size was deemed necessary. To validate the hypothesis, more than the available 6,254 cropped bounding boxes are required, leading to the training of an object detector to identify more unlabelled lesions in the data set.

3) Rationale for Additional Lesions: Several factors indicated the presence of more unlabelled lesions in the dataset: 1. The FFA-IR article mentioned that only 5% of cases were healthy, supporting the possibility of a higher number of lesions compared to the provided 6,254 bounding boxes. 2. The initial promise of 12,166 bounding boxes, as opposed to the delivered 6,254, further supported this belief.

4) Methodological Overview: Given the assumption about the presence of additional unannotated lesions in the dataset, the methodology is structured into three distinct parts. First, determine suitable cases for training the object detector. Next, proceed with the actual training of the object detector. Finally, crop the detected lesions from the images and use them to train the MRG model.

A. Part 1: Identification of Lesion-free Cases Using Frequent Words

1) Objective: For effective training of object detectors, both positive and negative cases are essential. The paper provided annotated cases that serve as the positive samples. To identify the negative samples, a method was developed to categorise images based on shared words in their medical reports. This strategy is based on the assumption that there are unannotated lesions in the dataset. By comparing lexical similarities with the 46 known lesion categories, the aim is to pinpoint potential unannotated lesion cases. Identifying these cases allows for a more accurate representation of negative samples when training the object detector, complementing the provided positive samples.

2) Procedure: **Word Frequency Analysis:** The preliminary step involved identifying the most common words for each category. Stoppage words were removed and through an iterative process, various word counts were tested, ranging from 10 to 23. The number 14 emerged as the optimal choice, not arbitrarily, but based on its performance in the category matching phase. Specifically, when using 14 words, there was a harmonious balance achieved: it provided sufficient specificity to the category while avoiding the pitfalls of both underfitting and overfitting. **Category Matching:** Two metrics were calculated. First, the proportion of cases within a category sharing the 14 common words was determined. A 70% threshold was set, meaning reports had to share 70% of these common words to be considered a match. Second, the commonality of these words with other medical reports was assessed.

Case Filtering: Unlabelled cases that exhibited word commonalities with any lesion category were filtered out. Only those cases displaying no lexical similarities were retained as negative samples for training the object detector.

B. Part 2: Object Detector Training

1) Objective: The primary aim is to train an object detector using the labeled data combined with the lesion-free cases identified in Part 1.

2) Procedure: **Data splitting:** A stratified split across the categories was employed to guarantee an equitable distribution of each category between training and testing sets. The data set was partitioned into 80% for training and 20% for testing.

Case Selection: From the 269 cases that encompassed 6,254 images, a corresponding number of cases were randomly selected from those pinpointed in Part 1, which did not align with any lesion categories. From these, ten images were chosen from each case, summing up to 2,683 images. This distribution of positive and negative images was initially chosen because the software used at initially restricted the dataset size for training an object detector to 9,000 images. While a different methodology was eventually adopted, this dataset size had already been set. For future endeavours, it might be beneficial to modify the dataset to mirror a 50/50 split of annotated and unannotated cases that were identified as healthy.

Model Training: The selected cases were processed using the Faster-RCNN model, implemented via Detectron2. The initial training spanned 20,000 iterations with a learning rate of 0.0025 and a batch size of 2 images. Upon evaluation, adjustments were made: the iterations were reduced to 15,000, the learning rate was increased to 0.005, and the batch size was set at 4 images, with solver steps at 10,000 and 12,500.

Performance Evaluation: Following the training, the model's effectiveness was gauged using the test set, with precision as the primary metric. While mean average precision (which balances both precision and recall) is commonly employed, precision was specifically chosen for this study. The rationale behind this choice is the primary interest is in ensuring that when an image is identified, it accurately reflects the lesion in question, rather than ensuring all images with lesions are identified.

Scope of Lesion Detection: Due to data constraints, the model was trained to detect only the four most common lesions. While there are various lesions that could be identified, the limited data available necessitated this focus to ensure the accuracy and reliability of the model's detections.

C. Part 3: Image Cropping and MRG Evaluation

1) Objective: The aim is to crop the detected lesions and subsequently assess the improvements in Medical Report Generation (MRG).

2) Procedure:: Given the diversity in image sizes within the FFA-IR data set, with 13 distinct sizes seen in figure 3.1 (appendix), a cropping strategy was employed. When an image was cropped, it was scaled up to the nearest larger size available in the data set's size dictionary. To maintain image quality and consistency with other data set images that required scaling, the peripheries of the up scaled images were padded with black pixels. This approach mirrored the treatment of other FFA-IR images, where black pixels had already been added to compensate for lost details and to conform to one of the data set's 13 standardised sizes.

Following the cropping process, these images were integrated with the un-cropped images in the training set, forming the data set used to train the model.

The model's performance was then evaluated using both the validation and test sets. Evaluation metrics included BLEU_1, BLEU_2, BLEU_3, BLEU_4, METEOR, CIDEr, and ROUGE_L. The baseline model was trained using the parameters and code provided by the authors of the paper [7]. These were the baseline metrics used for comparison in this investigation.

D. Conclusion of Methodology

In summary, this methodology was designed to address the challenges posed by the dataset's limitations and to harness the potential of object detection in enhancing the medical report generation process. The multi-faceted approach, from leveraging lexical similarities in medical reports to the strategic cropping of images, underscores the depth and breadth of the research. The subsequent chapters will outline the results and discussions, demonstrating the efficacy of the proposed methods and their implications in the broader context of medical imaging and diagnostics.

VI. Results

A. Introduction

This chapter presents the findings of the research, systematically divided into three pivotal sections. Each section corresponds to a distinct phase of the study:

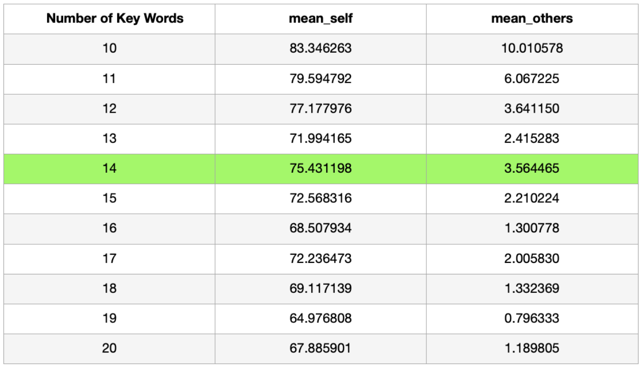
B. Part 1: Identification of lesion-free cases using frequent words

1) Objective and Method Overview::

The method consisted of testing a number of words from 10 to 23 to determine the optimal number of key words that maximised recall and minimised the false positive rate. Analysing the category matching results of the true positives for each category (mean_self) and the false positive rate for each category (mean_others), the results can be seen in Figure 1.1. A threshold of 70% was employed, meaning in order for a case to be deemed as a positive prediction, it only needed to contain 70% of the most common words.

2) Selection of Common Words:

Figure 1.1 displaying the relationship between the number of key words selected and their corresponding mean values for both the target category ("mean_self") and other categories ("mean_others").|



Number of Key Words	mean_self	mean_others
10	83.346263	10.010578
11	79.594792	6.067225
12	77.177976	3.641150
13	71.994165	2.415283
14	75.431198	3.564465
15	72.568316	2.210224
16	68.507934	1.300778
17	72.236473	2.005830
18	69.117139	1.332369
19	64.976808	0.796333
20	67.885901	1.189805

From the table, it's evident that selecting 14 key words yielded a favourable balance with a high "mean_self" value and a relatively low "mean_others" value, indicating a low false positive rate. While 12 key words also produced promising results, the choice of 14 was selected by its performance in category 40. Specifically, the 12-key word set failed to identify any matches with the common words in category 40 (figure 1.2.1), whereas the 14-key word set identified them at a rate of 25% (figure 1.2.2). This distinction underscores the importance of selecting an optimal number of key words to ensure comprehensive and accurate categorisation.

Figure 1.2.1:
12 Key Words Results: A bar graph depicting the proportion of true positives (recall) for each category in blue and the false positive rate for each category in red.

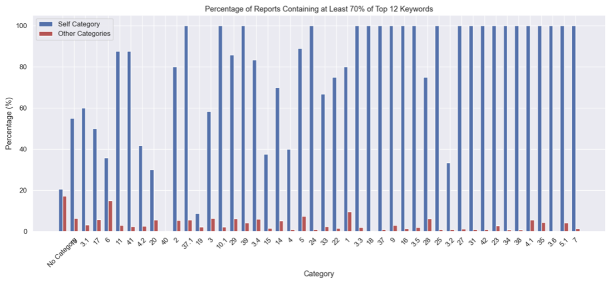
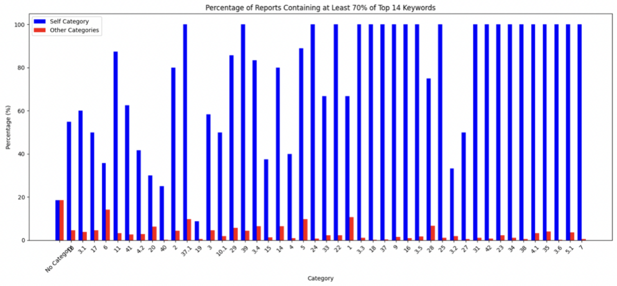


Figure 1.2.2:
14 Key Words Results. A bar graph depicting the proportion of true positives (recall) for each category in blue and the false positive rate for each category in red.



C. Category Matching Results

The results presented in figure 1.2.2 are notably impressive. The blue bars represent the recall, indicating the proportion of true positives identified for each category. On average, across all categories, the recall stands at a commendable 75.43

Conversely, the red bars represent the false positive rate, which measures the proportion of cases incorrectly identified as belonging to a category. The average false positive rate across all categories is a mere 3.56%. These false positives are particularly significant as they indicate cases that show similarities to positive cases, even though they have not been initially labelled as too belonging to the category. As part of the data refinement process, these false positives will be excluded from the data set when identifying negative cases for training the object detector.

1) Correlation Analysis:

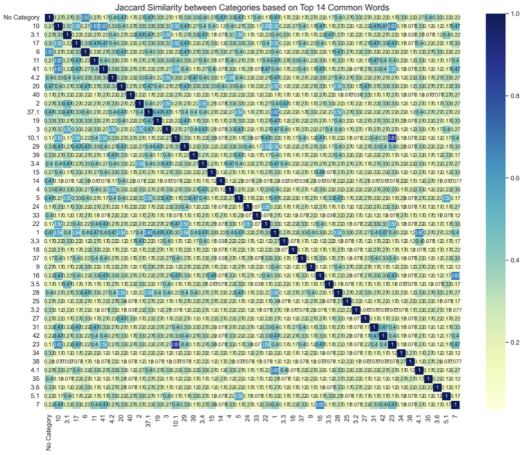


Figure 1.3: A heat-map illustrating the Jaccard similarity coefficients between categories based on their top 14 common words.

The heat-map in figure 1.3 provides insights into the degree of similarity between categories based on their shared common words. A predominant yellow hue across the squares indicates a low degree of similarity. This suggests that the top 14 common words in one category are often distinct from those in other categories. This low degree of similarity reaffirms the efficacy of the method, which leverages common words in reports to categorise and differentiate between various categories.

2) Threshold selection: A threshold of 70% was set as the criterion for category matching. While the possibility of exploring various thresholds was considered, the expansive scope of this investigation meant that evaluating multiple thresholds for each keyword count would have significantly expanded the testing parameters. Given the existing results, which identified cases containing lesion categories to a reasonably high accuracy, further exploration was deemed unnecessary for this study. However, future research should certainly contemplate testing a broader range of threshold values. The decision to limit the threshold testing was primarily a result of time constraints.

3) Summary: This section embarked on a methodological journey to categorise images based on shared words in their medical reports. The key findings and takeaways from this section are:

Optimal Word Selection: Through rigorous testing, 14 was identified as the optimal number of key words to use for categorisation. This number not only provided a high recall for most categories but also ensured a low false positive rate, striking a balance between specificity and sensitivity.

Correlation Analysis: The heat-map analysis, based on the Jaccard similarity coefficients, revealed a low degree of similarity between categories. This indicates that the method of using the top 14 common words is effective in distinguishing between categories, ensuring that each category is unique in its lexical makeup.

In essence, the methodology employed in Part 1 has proven to be both effective and efficient in categorising images. By leveraging shared words in medical reports, we've established a robust system that accurately identifies and differentiates between various lesion categories.

D. Part 2: Object Detector Training and Evaluation

1) Data Distribution and Split: Based on the methodology's specified parameters for training the object detector, the model was set to run for 15,000 iterations with a learning rate of 0.005. The batch size was configured at 4 images, and solver steps were designated at 10,000 and 12,500 intervals. Figures 2.0.1, 2.0.2, and 2.0.3 in the appendix showcase the progression of training accuracy and training loss. Given more time, a broader exploration of hyper-parameter combinations would have been advantageous. However, given the time constraints, this model represents the best outcome achieved within the available time frame. The model appears to cut off iterations right when the training loss and accuracy are about to plateau which is an ideal observation as it minimises the risk of a model that overfits.

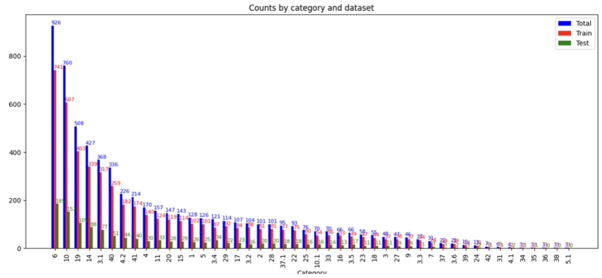
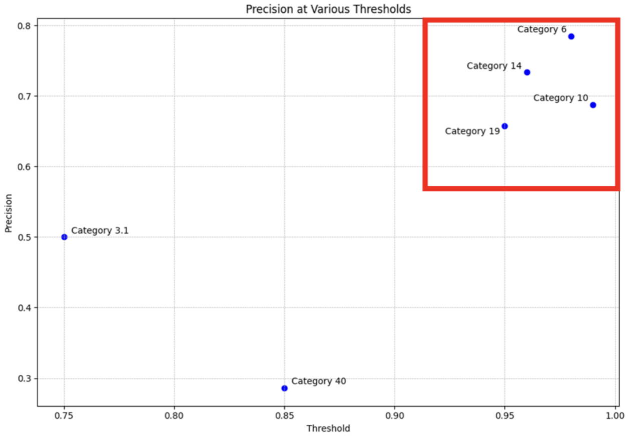


Figure 2.1: A stacked bar graph illustrating the distribution of samples across different categories. The blue bars represent the total count of annotated images for each category, the red bars denote the 80% training split, and the green bars indicate the 20% testing split.

The graph in figure 2.1 highlights the imbalance in the distribution of samples across categories. Categories such as 5.1, 38, and 36 have only 3 bounding boxes, making it challenging to train an object detector effectively for these lesions. In stark contrast, categories 6, 10, 19, and 14 boast over 400 samples each, showcasing a significant disparity in data distribution. Due to this imbalance, certain categories with sparse data were deemed unsuitable for accurate training and detection. The implications of this decision are further elaborated upon in the subsequent section.

2) Precision Analysis:



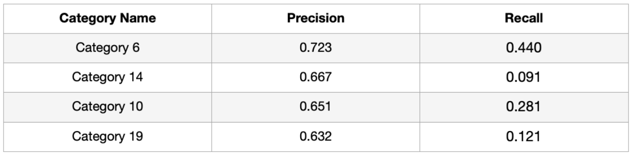
- **Figure 2.2:** A scatter plot showing the precision scores for the 6 most populous categories, highlighting the top four categories in red.

Figure 2.2, added the six most populated categories and looked at their performance on the testing set. A range of confidence threshold values from 0 to 1 were tested, incrementing by 0.01 each time. The results show the highest precision obtained for the top 6 categories.

Using this precision analysis, it reaffirms the observations made in the data distribution section. Categories 6, 14, 10, and 19, being the most populous in the data set, also exhibit the highest precision scores, ranging between 0.65 to 0.8. This performance juxtaposes the subsequent categories, 3.1 and 40, which show a noticeable drop in precision. Given these results, the decision was made to exclude the lesser-populated categories from further investigation and only focus on the top four categories. The rationale behind this choice is the insufficient data available for the other categories, which would likely lead to inconclusive or unreliable results due to the lack of accuracy in detecting lesions.

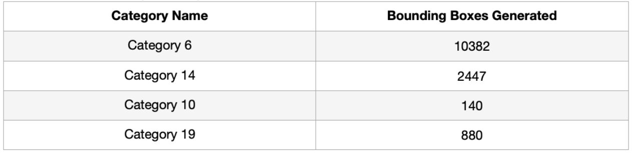
3) Implementing the confidence threshold into the predictions of the object detector. : When testing the object detector initially a high confidence threshold of 0.97 was set to ensure similar precision values to that seen in the graph above. However, this high threshold resulted in a limited number of predictions, see table 2.1.1, and this is due to the low recall values, see figure 2.1.2.

Figure 2.1.2 Precision and Recall for a confidence threshold of 0.97



Category Name	Precision	Recall
Category 6	0.723	0.440
Category 14	0.667	0.091
Category 10	0.651	0.281
Category 19	0.632	0.121

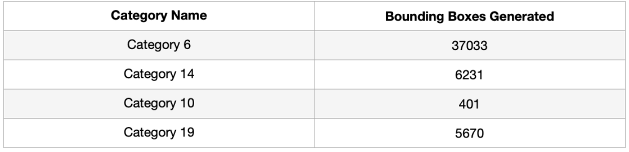
Figure 2.1.1 Bounding boxes generated for 0.97 confidence threshold



Category Name	Bounding Boxes Generated
Category 6	10382
Category 14	2447
Category 10	140
Category 19	880

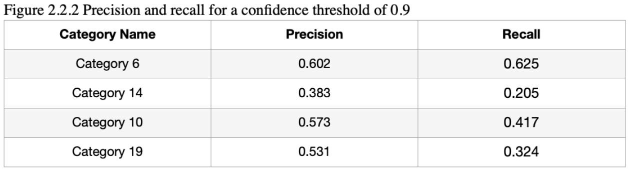
The predictions in table 2.1.1 represent 1.98% of the training set which is not desirable. This limitation in predictions is caused by the limited labelled data provided by the other categories. The initial plan involved detecting 46 lesions so a high precision was deemed appropriate due to the larger number of lesions being detected. However now that only 4 categories are being detected there were less predictions than initially anticipated. Since this threshold of 0.97, didnt form enough predictions, precision was sacrificed and recall was increased by using a confidence threshold of 0.90

Figure 2.2.1 Bounding boxes generated for a 0.9 confidence threshold



Category Name	Bounding Boxes Generated
Category 6	37033
Category 14	6231
Category 10	401
Category 19	5670

Figure 2.2.2 Precision and recall for a confidence threshold of 0.9



Category Name	Precision	Recall
Category 6	0.602	0.625
Category 14	0.383	0.205
Category 10	0.573	0.417
Category 19	0.531	0.324

This change increased the number of predictions, see figure 2.2.1, which now represented 5.94% of the training set. The trade off of precision for recall can be seen in figure 2.2.2, the results show that the precision for category 6 is roughly 60%, categories 10 and 19 are 57.3% and 53.1% and finally category 14 has a very low precision of approximately 38.3%.

These precision values, while not desired, were the best obtained while also maximising the number of predictions. Lowering the confidence threshold anymore caused precision to drop too much and increasing the threshold resulted in not enough data being generated. To combat this issue an alternative approach was explored.

This approach consisted of retraining the object detection model where it grouped all 46 lesions into a single category. This model, unfortunately, under performed, achieving a mean average precision score accuracy of 0.105 which under performs compared to the model detecting the 4 categories at a 90% confidence threshold. The poor performance can be attributed to the inherent differences between lesion types, which the model struggled to generalise.

4) Summary: The primary challenge faced during this phase was the discrepancy between the expected and actual data availability. The data set, as promised in the FFA-IR paper, was expected to offer nearly twice the number of labelled lesions. Such a comprehensive data set would have potentially facilitated the training of a more precise object detector, capable of identifying more than just four lesions and would improve the accuracy in predictions.

E. Part 3: Image Cropping and MRG Evaluation

1) Image Processing: In this phase, the images underwent a systematic preprocessing routine using the Faster-RCNN object detector trained in part 2. Once an object was identified, it was resized to align with one of the 13 predefined dimensions, as depicted in figure 3.1. This resizing strategy was crucial to ensure uniformity across the dataset. To further maintain consistency, especially with FFA images that lacked certain information, the images were resized upwards to the nearest predefined size. Any additional space was padded with black pixels, ensuring that the core information remained central and undistorted. This method stayed consistent with how the dataset was already resizing images to one of the 13 dimensions.

2) MRG Model Performance:

Figure 3.2 Performance of proposed model vs baseline

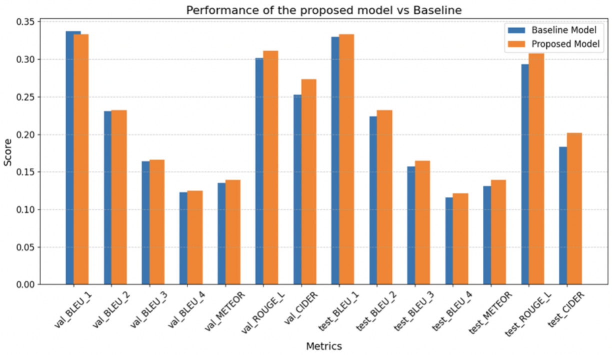
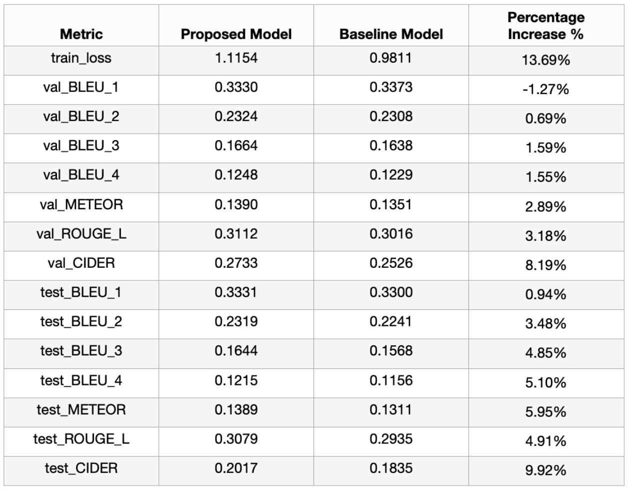


Figure 3.3 Table of results of proposed model against baseline model



Metric	Proposed Model	Baseline Model	Percentage Increase %
train_loss	1.1154	0.9811	13.69%
val_BLEU_1	0.3330	0.3373	-1.27%
val_BLEU_2	0.2324	0.2308	0.69%
val_BLEU_3	0.1664	0.1638	1.59%
val_BLEU_4	0.1248	0.1229	1.55%
val_METEOR	0.1390	0.1351	2.89%
val_ROUGE_L	0.3112	0.3016	3.18%
val_CIDER	0.2733	0.2526	8.19%
test_BLEU_1	0.3331	0.3300	0.94%
test_BLEU_2	0.2319	0.2241	3.48%
test_BLEU_3	0.1644	0.1568	4.85%
test_BLEU_4	0.1215	0.1156	5.10%
test_METEOR	0.1389	0.1311	5.95%
test_ROUGE_L	0.3079	0.2935	4.91%
test_CIDER	0.2017	0.1835	9.92%

The preprocessing of the images seemed to enhance the accuracy of the MRG model, as seen in figure 3.2 and figure 3.3. The validation metrics predominantly showcased positive shifts, with the exception of BLEU1 which experienced a marginal dip of 1%. The most commendable improvement was observed in val_Cider, registering an 8.2% increase, moving from a baseline of 0.252 to 0.273 in the refined approach.

The testing metrics further echoed this trend of improvement. Every category marked an improvement, with BLEU_1 showing the lowest rise of the metrics tested with a modest rise of 0.95%. In contrast, the metric CIDER stood out, reflecting a remarkable 9.9% enhancement in testing accuracy. Metrics BLEU_4, METEOR, and ROUGE_L all experienced an increase close to 5% in the testing set.

It is worth noting that the proposed model did register a slightly elevated training loss of 1.11, in comparison to the baseline's 0.98. This indicates that the model didn't fit the training data as well as the baseline. While a higher training loss might initially seem concerning, it can often be indicative of a model that generalises better to unseen data. In this context, the slightly higher training loss might suggest that the proposed model, while not fitting the training data as tightly, might be more robust and versatile when encountering new data sets.

3) Comparing Reports: Figure 3.4, found in the appendix, showcases select generated reports. The initial three represent instances where the proposed model significantly surpassed the baseline in terms of BLEU score. Conversely, the subsequent three highlight instances where the baseline outperformed the proposed model. Although meaningful conclusions cannot be solely derived from these isolated cases, it's evident that there are specific instances where one model distinctly outperforms the other. This showcases that the proposed model doesnt surpass the baseline with every case.

VII. Discussion

The results found that the proposed MRG model yielded a higher accuracy across all metrics in the testing set compared to the baseline model. The purpose of this discussion is to interpret the key findings, compare to the literature, discuss the implications, highlight the limitations, and finally recommend future directions for all three areas of this investigation.

A. Interpretation of Key Findings

1) Categorisation using Common Words Evaluation: The use of common words in medical reports as a categorisation tool for images has proven to be a significant methodological advancement. The approach centred on harnessing the top 14 words associated with each lesion category. From here, it proved to be an effective method to categorise cases based on the recurrence of these words in their reports. With a recall rate of 75.43%, the method not only showcased its accuracy but also highlighted a robust correlation between specific terminologies and their respective medical reports. The minimal overlap of these common words with other lesion categories, as evidenced by the heat map in figure 1.3, likely contributed to the impressively low false positive rate of 3.56%.

2) Object Detector Training and Evaluation: The challenges posed by an unbalanced data set were evident in efforts to train the object detector. Given the data's skewness, achieving robust detection across all 46 lesion categories proved elusive. Because of this the final object detector was only trained to detect 4 of the 46 categories of lesions. When using the object detector at a confidence threshold of 90% the results indicated a modest precision, slightly above 50%, for three of the four categories, figure 2.2.2, but a model with a higher precision was definitely desired.

When analysing the detections made by the model figure 2.2.1, category 6 dominated with 37,033 bounding boxes in the FFA-IR data set and a precision of 0.523. Category 14 had the least amount of bounding boxes with a total of 401 and the lowest precision of the four categories being 0.383. This is ideal as the category with the lowest precision made the fewest predictions lowering the number of false predictions in the data set.

The data's limitations were clear, since only 4 of the 46 categories had enough data to train the object detector, only 5.94% (41676/701693) of the training class consisted of detected objects. This paired with the low precision in predictions made the object detector the biggest area of concern in this research.

It is worth noting that the implications of reduced precision might be less severe than initially presumed. The reason for this belief is that most cases in the data set contain hundreds or even thousands of images, and so a few misidentified magnified regions might not dramatically affect the overall accuracy, especially when no lesions are detected. This limited effect would arise because many images within a case are similar, essentially relaying the same information. Therefore, an occasional misidentification might not significantly alter the overall context. However, for images that actually show lesions, it is magnifying information that may not be in the other images and focuses on those areas increasing the chance for the MRG model to detect the lesion, as cropping helps spotlight the lesion and remove irrelevant details. The reason for this hypothesis is due to the evident improved accuracy of the final MRG model.

In this research, the final object detector trained, detected these four lesions: Macular pucker (or Epiretinal membrane), Choroidal neovascularization, High myopia choroidal neovascularization, and Lacquer crack pathological myopia macular lesion.

3) Image Cropping and MRG Evaluation: The implementation of the proposed method led to notable enhancements in the MRG metrics compared to the baseline. While the training loss was marginally higher at 1.11 compared to the baseline's 0.98, the model demonstrated significant improvements across all evaluation metrics in the testing set and all but one in the validation set. The most notable improvements are seen with the CIDEr metric with an increase of 9.9% in the testing accuracy.

Differential Sensitivity: While metrics like BLEU, METEOR, and ROUGE focus on specific lexical and phrasal overlaps, CIDEr places a strong emphasis on the consensus of descriptions. This makes it especially adept at evaluating the accuracy with which key image details are captured.

Harmonising with Human Perspectives: The significant improvement in the CIDEr score indicates that the model's generated reports align more closely with human interpretations, highlighting the most relevant and crucial details of the images.

4) summary: Overall, the results largely affirm the initial hypothesis that focusing on lesions and reducing extraneous noise can improve MRG accuracy. However, the extent to which the results support this is unclear as there is also the possibility that merely removing noise from a selection of images regardless of if they contain lesions or not will improve the MRG models accuracy. This is a result caused by the object detector's low precision, which led to the cropping of many regions without lesions.

B. Comparison with Existing Literature

1) Comparing Categorisation of Lesions to Common Text Mining Approaches: When comparing the findings of this research with existing literature for detecting lesions with word association, it's evident that this research differs from many of the existing studies. This research falls under the study of text mining used for medical classification. An example of a paper that used radiology reports to classify a medical condition can be seen in, Using Different Feature Extraction Methods to Identify Fractures of the Distal Fibula [35], which utilised free-text radiology reports about leg fractures, and used text mining techniques to identify cases of a distal fibula.

The mentioned study reported an impressive accuracy of 0.97 and an AUC of 0.98. Their most effective text representation method was the Bag-of-Words (BOW), and among classifiers, Neural Networks (NN) stood out with the highest AUC.

However, their research had a narrower focus, concentrating on classifying a specific type of leg fracture. In contrast, this study delved into 46 categories, and faced serious challenges in prediction due to insufficient data.

It's worth noting that in this context, the AUC metric isn't directly applicable since there is a lack of a clear understanding of what constitutes as a false positive since images are not labelled as negative cases. As a result, precision cannot be calculated. Adding to this point, the methods they employed such as Neural Networks and Bag-of-Words require the knowledge of true negative cases. This research was attempting to identify these true negative cases so adopting a similar methodology was not possible.

In a direct comparison, this method achieved a recall of approximately 75%, which, while commendable, is lower than the 0.97 accuracy reported in the aforementioned study. Given the data set's constraints, we've treated recall as a proxy for accuracy, as discerning between false positives and unlabelled true positives remains challenging.

Our approach, while distinct from mainstream literature, has demonstrated it's worth in identifying a broad spectrum of lesions. While integrating techniques like Bag-of-Words and Neural Networks could potentially enhance the methodology, the prerequisite would be a more comprehensive data set, inclusive of both lesion cases and confirmed non-lesion examples.

2) Literature Comparison in the Context of Object Detection: When examining the existing literature, two notable papers, [8] and [9], employed object detection to augment image captioning.

However, neither study disclosed the accuracy of their object detectors, leaving a gap in comparative analysis.

There were two other papers mentioned in the literature review which used object detection on Fundus images and they each used different object detectors.

The study [32] employed the Faster RCNN for object detection. Impressively, they pinpointed the macula, a central, oval area of the retina, with an accuracy of 0.811. This surpasses the best precision score of 0.790 for category 6 at a 0.98 confidence threshold. Given their advantage of having 2893 labeled macula images, it's conceivable that with similar data quantities, the object detector trained in this paper could have optimised to obtain a higher accuracy.

Continuing on from this, it's pertinent to mention that comparing precision to accuracy directly isn't the most ideal or equitable form of assessment. Precision focuses on the correctness of the positive predictions, while accuracy evaluates the proportion of all predictions that are correct. In the context of this study, the primary emphasis has been on precision, given the criticality of correctly identifying positive instances without introducing many false positives. Therefore, while juxtaposing the precision score with another study's accuracy might not present a completely analogous comparison, it's the closest and most relevant analysis available, considering the metrics prioritised in this research.

In the study [33], a distinctive convolutional neural network strategy was applied, demonstrating a marked accuracy of up to 91% and specificity of 100% in detecting and categorising diabetic retinopathy (DR) and its related clinical signs using the YOLOv5 algorithm. This approach, utilising a substantial dataset of 8,000 high-resolution fundus images and incorporating detailed annotations of DR's clinical signs, has shown appreciable results, particularly in the nuanced identification of specific DR indicators. When contrasted with this study, a notable divergence emerges in terms of methodological approach and dataset volume, wherein this study employed the Faster RCNN model and faced limitations regarding the volume of available annotated data.

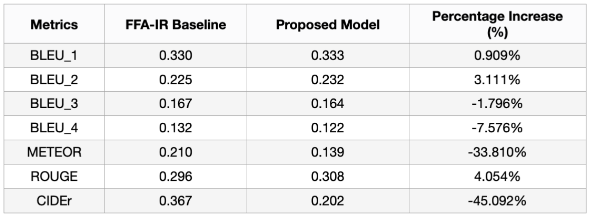
This papers primary focus was precision, achieving peak scores for the top four categories in the range from 0.65 to 0.8, to attempt to minimise false positives. While both studies leaned heavily into object detection methodologies, the disparity in annotated data availability, detection algorithms, and focal metrics (precision vs. accuracy) underscores the varied challenges and outcomes experienced. The comparison, therefore, provides valuable insights into the adaptability and outcomes of different object detection algorithms under varied constraints and objectives within the medical imaging domain.

3) Baseline Comparisons and Model Performance: In the FFA-IR paper, the baseline results they reported compared to what was obtained when running the same code using the same baseline parameters were somewhat different. Notably, their METEOR score was considerably lower, while their CIDEr score was reported to be much higher. Due to certain inconsistencies in their paper, such as missing annotations, removal of files, and the absence of responses to emails, the analysis of these results was taken with a degree of skepticism.

When aligning the results from the proposed model with the FFA-IR baseline [7], as shown in figure 2.3, a stark disparity emerges notably, the proposed model underperforms by 45% in the CIDEr metric and 33.8% in the METEOR metric. The authors need to clarify the parameters in their baseline model to ensure reproducibility across different platforms, as running the code they provided did not yield these initial results in the baseline.

It is essential to note that these discrepancies don't diminish the advancements this research has achieved over the baseline as the proposed model and the model presented in that paper use different parameters. Following on from this, it would be interesting to see if training the MRG model using their parameters and the dataset created in this experiment would cause the same improvements in accuracy.

Figure 2.3 The baseline in the FFA-IR baseline compared to the results of the proposed model in this paper.



Metrics	FFA-IR Baseline	Proposed Model	Percentage Increase (%)
BLEU_1	0.330	0.333	0.909%
BLEU_2	0.225	0.232	3.111%
BLEU_3	0.167	0.164	-1.796%
BLEU_4	0.132	0.122	-7.576%
METEOR	0.210	0.139	-33.810%
ROUGE	0.296	0.308	4.054%
CIDEr	0.367	0.202	-45.092%

C. *Implications*

The findings of this research have far-reaching implications for the realm of medical imaging and report generation. These implications extend not only to the technical aspects of the field but also to the broader spheres of clinical practice, patient care, and the overall healthcare system.

Categorising Lesions in Medical Imaging Through Common Words: The first segment of research has pioneered the use of text mining to classify lesions in the FFA-IR dataset. Specifically, it has successfully categorised lesions based on medical reports for 46 distinct types within the context of FFA images.

Potential to Enhance Medical Report Generation (MRG): Using the results of categorising lesions based on key words, there emerges the possibility of integrating this into MRG. By ensuring that an MRG report encompasses the 14 most prevalent words associated with diagnosing a lesion, the accuracy and reliability of these reports could be significantly enhanced.

Object Detection in Medical Imaging: The second part of this research introduced a specialised object detector. The detector, is adept at filtering results and pinpointing the four most prevalent labelled lesions in the data set with the precision of detecting these lesions sitting in the range between 0.65 and 0.8 depending on the lesion. Notably, this is the inaugural object detector trained using the FFA-IR dataset.

Advancements in Medical Report Generation (MRG):

The third segment of this study underscores a pivotal advancement in MRG. Through this research, a notable enhancement in accuracy has been demonstrated, providing tangible evidence that the methods of this experiment can significantly elevate the quality and accuracy of medical report generation.

D. *Limitations and Constraints*

Every research endeavour faces its unique set of challenges, and this study is no exception.

1) Dataset Limitations: **Part 1:** A more substantial collection of annotated images, preferably validated by an ophthalmologist and encompassing confirmed images without lesions, would have established a stronger foundation. Such a dataset would have enabled a more comprehensive testing of categories and a more precise calculation of metrics, especially when discerning false positives with confidence. By doing this, other methods such as Bag of Words and Neural Networks could be applied to improve the accuracy of categorising cases.

Part 2: The dataset's limitations were again evident. A richer dataset would have facilitated the detection of a broader range of lesions meaning that a larger selection of images could be identified allowing a higher precision to be preserved. The current dataset's constraints necessitated a focus on only four out of the 46 lesions, leading to a compromise on precision to boost the number of predictions.

Part 3: The primary constraint in this segment was the object detector from Part 2, which was inherently limited by the available data. Additionally, the cropping methodology employed might not be the most optimal, suggesting room for refinement. This could be done by potentially altering the R2Gen code [10] to take an image of any size which is not necessarily limited to one of the 13 options in figure 3.1.

2) Methodological Constraints: In part 1 time constraints meant that only the threshold of 70% was investigated. A more exhaustive exploration might have yielded different insights and an increase in accuracy.

In part 2, the study could have benefited from testing a wider array of object detectors, such as YOLO and SSD. However, the extensive training time and the project's scope made this challenging. Furthermore, a more comprehensive exploration of hyperparameters, iterations, training loss, and images per batch would have been ideal. The potential for a validation set and experiments with k-fold stratified training was also left unexplored, primarily due to dataset limitations and the timeintensive nature of training multiple models. Finally testing the training of the object detector on different splits of images that contain lesions and no lesions could have lead to improvements.

In part 3, the evaluation metrics employed primarily gauge word similarity rather than the actual semantic value of the medical reports. Engaging an ophthalmologist to interpret and assess the underlying meaning of the model's outputs would yield a deeper and more accurate understanding of the results.

3) Assumptions and Their Implications:: Certain assumptions underpin this study, and it's crucial to understand how they might influence the interpretation of results. For instance, the assumption that the cases that did not share common words with categories were identified as lesion free images when this may not be the case. Another is the assumption that these evaluation metrics determine the better MRG model when this may not be true, there are certain nuances in medical terminology these metrics cannot detect.

In conclusion, while this study offers valuable insights, it's essential to view the findings in light of these limitations. Future research can build upon this foundation, addressing these constraints to push the boundaries of what's possible in medical imaging and report generation. The main limitation is the dataset and I implore the authors of the paper to review, annotate, and provide more lesion data as well as updating the baseline parameters so others can obtain the same results.

E. Recommendations and Future Directions

1) Part 1 Recommendations: **Refinement with Time:** Given more time, the model from this segment could be further refined and optimised. This could be done by testing different thresholds and potentially testing the performance on more data, if available.

Adoption of Proven Methods: Drawing inspiration from other studies, such as [35], could be beneficial. For instance, utilising a bag-of-words representation combined with a neural network might yield better results. The choice to not use these methods in this research was due to the uncertainties about the dataset. If they updated the dataset to contain verified lesion-free FFA images these methods can be investigated.

2) Part 2 Suggestions: **Rethinking Precision's Role:** While the object detector's performance was not optimal, it brought forth an intriguing proposition: perhaps the precision of detecting lesions isn't the sole determinant for improving MRG accuracy. To validate this idea, one could attempt to identify all lesions in the dataset using the model trained in this experiment and not filtering the images to only include the four categories with the highest precision. Even if the precision for some categories is considerably low, an improvement in accuracy might still be observed compared to the baseline due to noise reduction in images. This approach could offer a deeper comprehension of this study's conclusions and shed light on whether the initial hypothesisthat solely cropping lesions improves the MRG modelholds true.

3) Part 3 Future Directions:: **Image Processing Improvements:** A potential area of improvement lies in the image processing phase. Adjusting the MRG code to allow it to handle images of any dimension would remove the requirement for black borders during image cropping and would remove the black noise from being present in the images. Even padding the images to contain the surrounding information of the original image might be an interesting avenue of investigation although it would introduce some noise.

Data Augmentation: The most crucial enhancement would be obtaining more annotated data. Should an ophthalmologist, or even the authors of the paper, supply the additional 6,000 bounding boxes as indicated, it would greatly enhance all three stages of this investigation.

Robust Evaluation Metrics: To truly gauge the model's efficacy, more rigorous evaluation metrics could be employed. For instance, having ophthalmologists evaluate the results, similar to the evaluation process taken in the FFA-IR paper [7], a more accurate assessment of the model's real-world applicability can be determined.

Utilising text mined dictionaries: Implementing the dictionary sourced from text mining in part 1 as a criteria to enhance the generated medical reports. Implementing this to work with the decoder in the MRG model could cause the generated reports to increase in accuracy as they would be guaranteed to use the same words that a majority of the other reports in the same category contain.

In conclusion, while this study has laid a solid foundation, there's ample room for enhancement and exploration. Future research can capitalise on these recommendations, pushing research further in the realm of text mining, object detection, and medical report generation.

VIII. Conclusion

The discussion delved deep into the intricacies of medical imaging and report generation, uncovering several key findings and insights:

Part 1: underscored the novelty of employing text mining to classify lesions based on medical reports, specifically for the 46 different types of lesions in the context of FFA images. This approach not only showcased the potential of text mining as a preprocessing step in object detection for medical report generation but also hinted at its utility in enhancing the quality of the medical reports generated, using the words as a criteria for the MRG model.

Part 2: introduced an object detector tailored to identify the four most prevalent lesions in the labelled data with precision in a range of 0.65 to 0.8 for these categories. However, a pivotal revelation from this segment was the potential reevaluation of the role of precision in the overall scheme of things. Contrary to initial beliefs, the research suggests that the emphasis might not necessarily have to be on precision alone. The final model, which utilised both cropped lesion images and cropped non-lesion images, demonstrated enhanced performance in the MRG model. One such explanation could be that while cropping lesion-containing images is of high importance, cropping non-lesion images may not be as impactful. This is likely due to the substantial volume of similar images in each category, which ensures that essential information remains intact. Another explanation could be that cropping all images or a percentage of images regardless of what lesion they contain could result in a model which has a higher accuracy due to the robust nature of this new dataset.

Part 3: marked a significant stride in Medical Report Generation (MRG), showcasing an improved performance, thereby validating the proposed methodology. The potential implications of this study suggest that object detection should become a pre-processing step in MRG for FFA images, removing noise from images containing lesions has been proven to improve the accuracy of generated medical reports.

In conclusion, this research has made substantial contributions to the field of medical imaging and report generation. By challenging conventional beliefs, introducing innovative methodologies, and highlighting potential areas of improvement, this study paves the way for future research endeavours that can further refine and optimise the process of implementing object detection to improve the accuracy of MRG. The findings underscore the importance of a holistic approach, where precision, data quality, and innovative methodologies collectively drive advancements in the domain.

References

[1] Van Nynatten L, Gershon A. Radiology wait times: Impact on patient care and potential solutions. University of Western Ontario Medical Journal. 2017 Dec 3;86(2):65-6.

[2] Rabb MF, Burton TC, Schatz H, Yannuzzi LA. Fluorescein angiography of the fundus: a schematic approach to interpretation. Survey of ophthalmology. 1978 May 1;22(6):387-403.

[3] Eye Disease [Internet]. IDF Europe Site. [cited 2023 Oct 16]. Available from: https://idf.org/europe/life-with-diabetes/diabetes-related-complications/eye-disease/

[4] Vyawahare H, Shinde P. Age-related macular degeneration: Epidemiology, pathophysiology, diagnosis, and treatment. Cureus. 2022 Sep 26;14(9).

[5] Pavlopoulos J, Kougia V, Androutsopoulos I, Papamichail D. Diagnostic captioning: a survey. Knowledge and Information Systems. 2022 Jul;64(7):1691-722.

[6] Shamshad F, Khan S, Zamir SW, Khan MH, Hayat M, Khan FS, Fu H. Transformers in medical imaging: A survey. Medical Image Analysis. 2023 Apr 5:102802.

[7] Li M, Cai W, Liu R, Weng Y, Zhao X, Wang C, Chen X, Liu Z, Pan C, Li M, Liu Y. Ffa-ir: Towards an explainable and reliable medical report generation benchmark. InThirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2) 2021.

[8] Zeng X, Wen L, Liu B, Qi X. Deep learning for ultrasound image caption generation based on object detection. Neurocomputing. 2020 Jun 7;392:132-41.

[9] Yang Z, Zhang YJ, Rehman SU, Huang Y. Image captioning with object detection and localization. InImage and Graphics: 9th International Conference, ICIG 2017, Shanghai, China, September 13-15, 2017, Revised Selected Papers, Part II 9 2017 (pp. 109-118). Springer International Publishing.

[10] Chen Z, Song Y, Chang TH, Wan X. Generating radiology reports via memory-driven transformer. arXiv preprint arXiv:2010.16056. 2020 Oct 30.

[11] Varges S, Bieler H, Stede M, Faulstich LC, Irsig K, Atalla M. SemScribe: Natural Language Generation for Medical Reports. InLREC 2012 May (pp. 2674-2681).

[12] Schlegl T, Waldstein SM, Vogl WD, Schmidt-Erfurth U, Langs G. Predicting semantic descriptions from medical images with convolutional neural networks. InInformation Processing in Medical Imaging: 24th International Conference, IPMI 2015, Sabhal Mor Ostaig, Isle of Skye, UK, June 28-July 3, 2015, Proceedings 2015 Jun 23 (pp. 437-448). Cham: Springer International Publishing.

[13] Kisilev P, Sason E, Barkan E, Hashoul S. Medical image description using multi-task-loss CNN. InDeep Learning and Data Labeling for Medical Applications: First International Workshop, LABELS 2016, and Second International Workshop, DLMIA 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 21, 2016, Proceedings 1 2016 (pp. 121-129). Springer International Publishing.

[14] Cho K, Van Merrinboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078. 2014 Jun 3.

[15] Wu L, Wan C, Wu Y, Liu J. Generative caption for diabetic retinopathy images. In2017 International conference on security, pattern analysis, and cybernetics (SPAC) 2017 Dec 15 (pp. 515-519). IEEE.

[16] HochreiterS S. Longshort-termmemory. NeuralComput9 (8): 17351780.

[17] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser , Polosukhin I. Attention is all you need. Advances in neural information processing systems. 2017;30.

[18] Li M, Liu R, Wang F, Chang X, Liang X. Auxiliary signal-guided knowledge encoder-decoder for medical report generation. World Wide Web. 2023 Jan;26(1):253-70.

[19] Li Y, Liang X, Hu Z, Xing EP. Hybrid retrieval-generation reinforced agent for medical image report generation. Advances in neural information processing systems. 2018;31.

[20] Li CY, Liang X, Hu Z, Xing EP. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. InProceedings of the AAAI Conference on Artificial Intelligence 2019 Jul 17 (Vol. 33, No. 01, pp. 6666-6673).

[21] Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: a method for automatic evaluation of machine translation. InProceedings of the 40th annual meeting of the Association for Computational Linguistics 2002 Jul (pp. 311-318).

[22] Vedantam R, Lawrence Zitnick C, Parikh D. Cider: Consensus-based image description evaluation. InProceedings of the IEEE conference on computer vision and pattern recognition 2015 (pp. 4566-4575).

[23] Banerjee S, Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. InProceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization 2005 Jun (pp. 65-72).

[24] Lin CY. Rouge: A package for automatic evaluation of summaries. InText summarization branches out 2004 Jul (pp. 74-81).

[25] Du L, Zhang R, Wang X. Overview of two-stage object detection algorithms. InJournal of Physics: Conference Series 2020 May 1 (Vol. 1544, No. 1, p. 012033). IOP Publishing.

[26] Sanchez SA, Romero HJ, Morales AD. A review: Comparison of performance metrics of pretrained models for object detection using the TensorFlow framework. InIOP Conference Series: Materials Science and Engineering 2020 May 1 (Vol. 844, No. 1, p. 012024). IOP Publishing.

[27] Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierar-

chies for accurate object detection and semantic segmentation. InProceedings of the IEEE conference on computer vision and pattern recognition 2014 (pp. 580-587).

[28] Girshick R. Fast r-cnn. InProceedings of the IEEE international conference on computer vision 2015 (pp. 1440-1448).

[29] Ren S, He K, Girshick R, Sun J. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems. 2015;28.

[30] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC. Ssd: Single shot multibox detector. InComputer VisionECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 1114, 2016, Proceedings, Part I 14 2016 (pp. 21-37). Springer International Publishing.

[31] Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. InProceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 779-788).

[32] Zhang Y, Ye X, Wu W, Luo Y, Chen M, Du Y, Wen Y, Song H, Liu Y, Zhang G, Wang L. Morphological Rule-Constrained Object Detection of Key Structures in Infant Fundus Image. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2023 Jan 5.

[33] Ramesh PV, Ramesh SV, Subramanian T, Ray P, Devadas AK, Ansar SM, Rajasekaran R, Parthasarathi S. Customised artificial intelligence toolbox for detecting diabetic retinopathy with confocal truecolor fundus images using object detection methods. TNOA Journal of Ophthalmic Science and Research. 2023 Jan 1;61(1):57-66.

[34] Abhishek AV, Kotni S. Detectron2 object detection & manipulating images using cartoonization. Int. J. Eng. Res. Technol.(IJERT). 2021;10.

[35] Dewald CL, Balandis A, Becker LS, Hinrichs JB, von Falck C, Wacker FK, Laser H, Gerbel S, Winther HB, Apfel-Starke J. Automated Classification of Free-Text Radiology Reports: Using Different Feature Extraction Methods to Identify Fractures of the Distal Fibula. InRFo-Fortschritte auf dem Gebiet der Rntgenstrahlen und der bildgebenden Verfahren 2023 May 9. Georg Thieme Verlag KG.

[36] Mustafa A, Rahimi Azghadi M. Automated machine learning for healthcare and clinical notes analysis. Computers. 2021 Feb 22;10(2):24.

IX. Appendix

A. Figure 3.1

Dictionary of cropping sizes (768, 768), (3180, 2696), (512, 512), (384, 384), (3200, 2600), (3180, 2600), (1600, 1200), (2448, 1956), (2285, 1900), (768, 576), (2124, 2056), (1444, 1444), (1536, 1536), (1024, 1024), (2285, 1880), (2260, 1880), (3216, 2136), (1111, 1438), (1130, 1038), (1386, 1105), (1074, 1113)

B. Figure 1.0.1

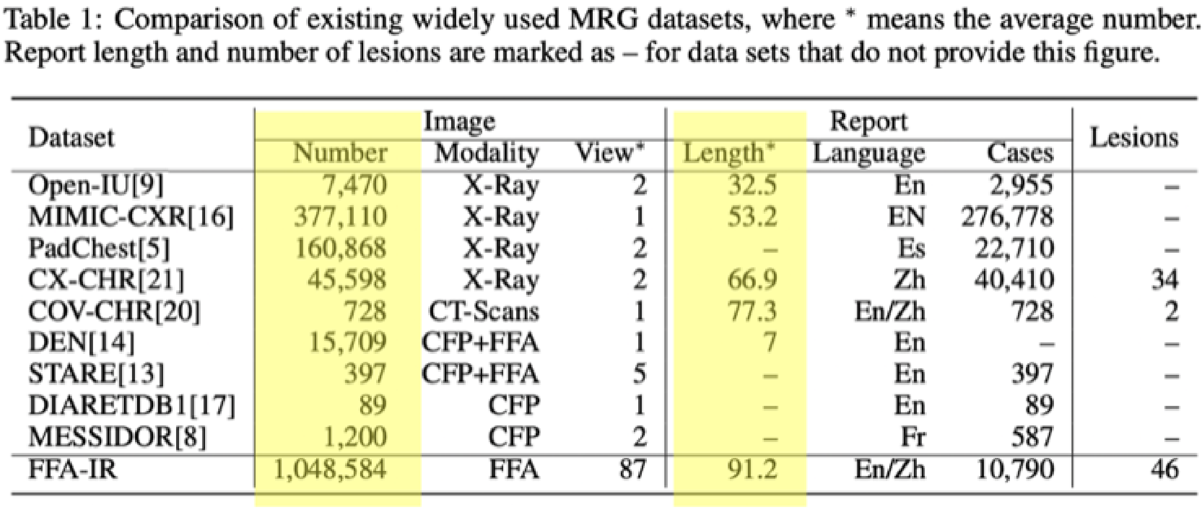
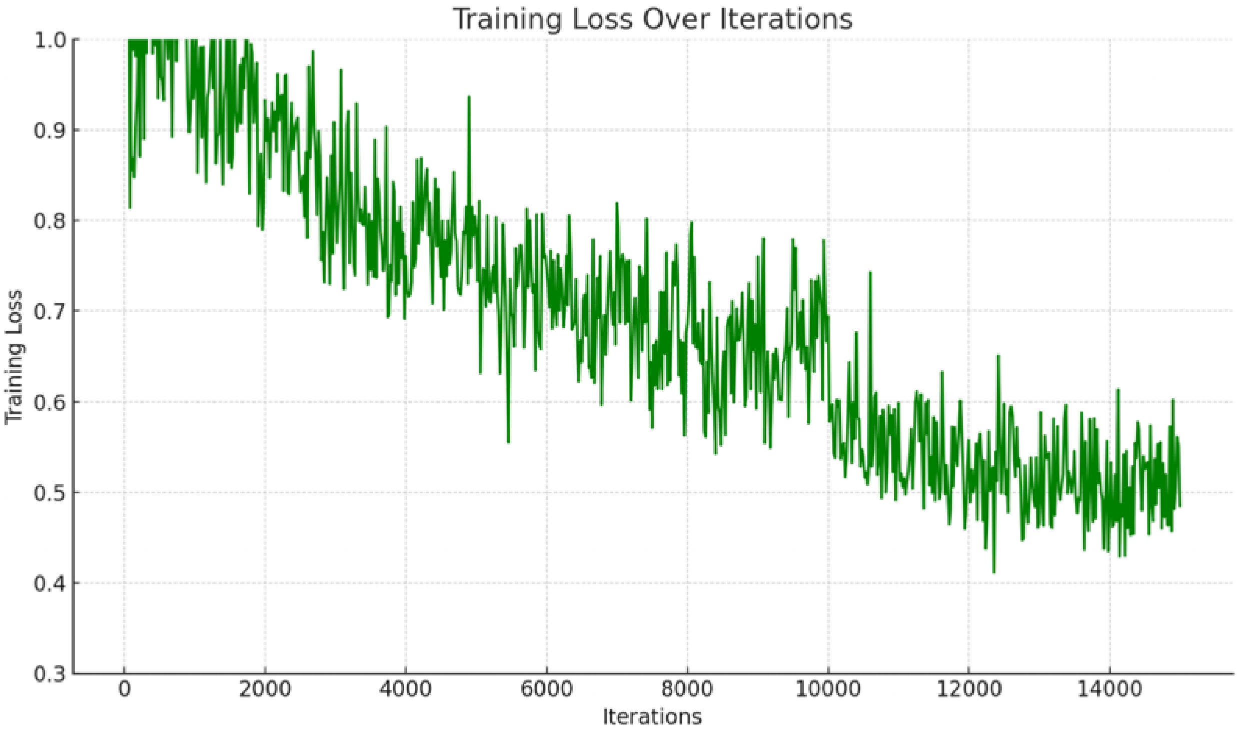


Table 1: Comparison of existing widely used MRG datasets, where * means the average number. Report length and number of lesions are marked as – for data sets that do not provide this figure.

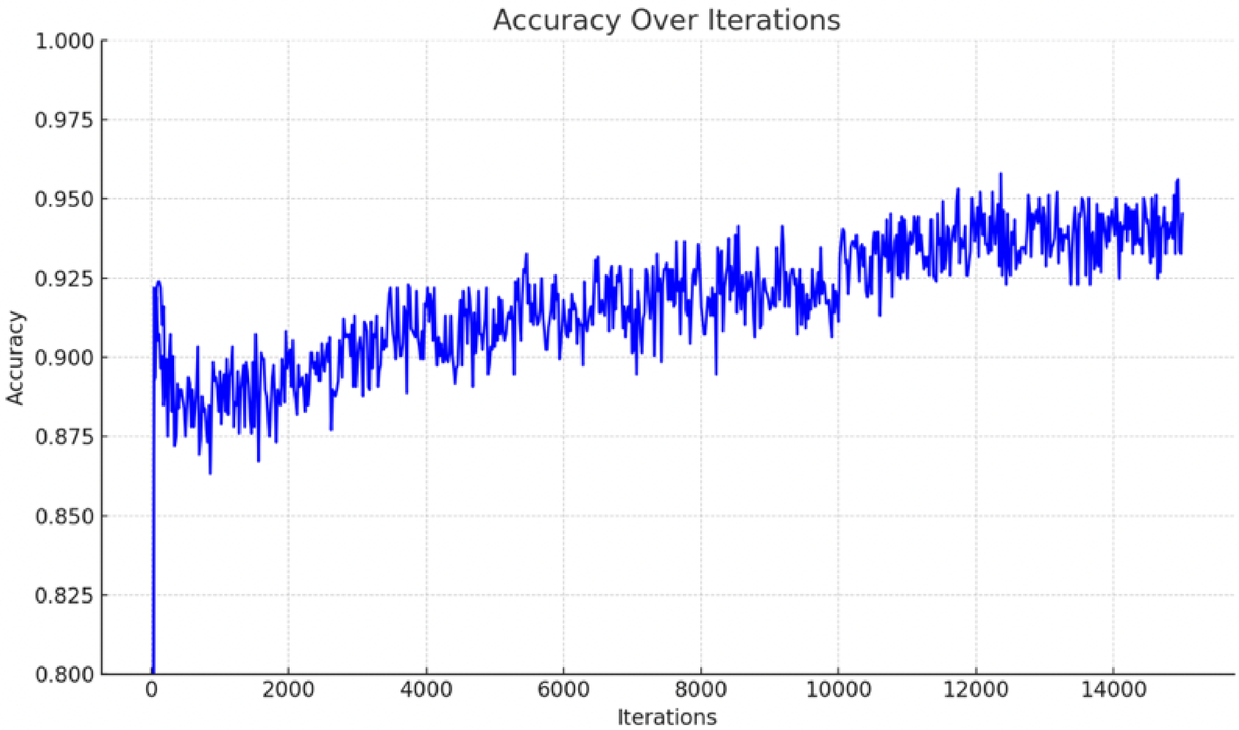
Dataset	Image			Report			Lesions
	Number	Modality	View*	Length*	Language	Cases	
Open-IU[9]	7,470	X-Ray	2	32.5	En	2,955	–
MIMIC-CXR[16]	377,110	X-Ray	1	53.2	EN	276,778	–
PadChest[5]	160,868	X-Ray	2	–	Es	22,710	–
CX-CHR[21]	45,598	X-Ray	2	66.9	Zh	40,410	34
COV-CHR[20]	728	CT-Scans	1	77.3	En/Zh	728	2
DEN[14]	15,709	CFP+FFA	1	7	En	–	–
STARE[13]	397	CFP+FFA	5	–	En	397	–
DIARETDB1[17]	89	CFP	1	–	En	89	–
MESSIDOR[8]	1,200	CFP	2	–	Fr	587	–
FFA-IR	1,048,584	FFA	87	91.2	En/Zh	10,790	46

C. Figure 2.0.1 to Figure 2.0.3

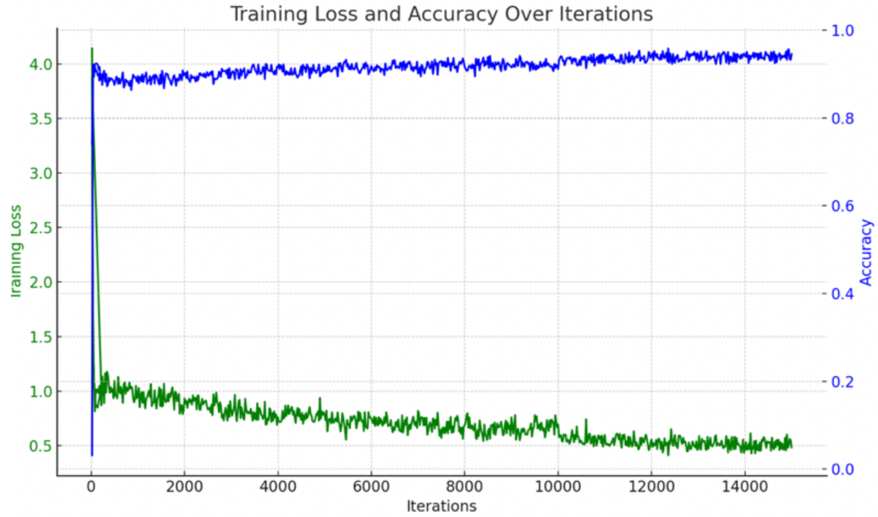
2.0.1



2.0.2



2.0.3



Training Loss and Accuracy Over Iterations
Training Loss
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
0
2000
4000
6000
8000
10000
12000
14000
Iterations

D. Figure 3.1

The three reports were the proposed model outperformed the baseline the most using the BLEU metric.

1) Proposed Outperforming Baseline Examples:

Case 1:

Ground Truth: No significant abnormal fluorescence was seen in all phases of fluorescein fundus angiography angiography in the left eye.

Baseline Model Output (BLEU Score: 0.3584): Fluorescein fundus angiography of the left eye showed scattered small dots of hyperfluorescence in the posterior pole and periphery during the angiography process with fluorescein leakage gradually with the prolongation of the angiography time.

Proposed Model Output (BLEU Score: 1.0): No significant abnormal fluorescence was seen in all phases of fluorescein fundus angiography angiography in the left eye.

Case 2:

Ground Truth: No significant abnormal fluorescence was seen in the macular retina. No significant abnormal fluorescence was seen in the mid-peripheral and peripheral retina. No significant abnormal fluorescence was seen in the optic disc. Baseline Model Output (BLEU Score: 0.3203): FFA of the left eye the fluorescence image of the optic disc is approximately normal. The retinal arteries are thin the veins are tortuous and dilated some of the walls are fluorescently stained the capillaries of the small retinal veins are dilated with increased permeability a large number of laser spots can be seen from the mid-periphery to the periphery and several vascular occlusion areas can be seen in the mid-periphery. Advanced edema in the macula.

Proposed Model Output (BLEU Score: 0.9434): No significant abnormal fluorescence was seen in the macula. No significant abnormal fluorescence was seen in the mid-peripheral and peripheral parts of the retina. No significant abnormal fluorescence was seen in the optic disc.

Case 3:

Ground Truth: FFA of the left eye preoptic disc neovascularization late fluorescent staining of the optic disc with clear borders. Massive microaneurysm from the posterior pole of the retina to the mid-periphery large vascular non-perfused area in the mid-periphery several pieces of neovascularization. Capillary dilatation with leakage in the macula late cystoid edema.

Baseline Model Output (BLEU Score: 0.2877): No significant abnormal fluorescence was seen in the macula. No significant abnormal fluorescence was seen in the mid-peripheral and peripheral parts of the retina. No significant abnormal fluorescence was seen in the optic disc.

Proposed Model Output (BLEU Score: 0.7465): FFA of the right eye no abnormal fluorescence of the optic disc with clear borders. Numerous microaneurysms from the posterior pole of the retina to the mid-periphery several vascular non-perfused areas and IRMA visible in the mid-periphery of each quadrant. Dilated capillaries with leakage in the macula late diffuse hyperfluorescent staining. In each of these cases, the proposed model's output is much closer to the ground truth compared to the baseline model's output, as evidenced by the significantly higher BLEU scores.

The three reports were the baseline model outperformed the proposed model the most using the BLEU metric.

2) Baseline Outperforming Proposed Examples:: **Case 1:**

Ground Truth: ”Fluorescein fundus angiography of the left eye showed small dots of hyperfluorescence in the posterior pole and periphery at the beginning of the angiogram with some slight fluorescein leakage gradually with the prolongation of the angiogram and lamellar blocked fluorescence in the periphery below.

Baseline Model Output (BLEU Score: 0.713): ”Fluorescein fundus angiography of the left eye showed scattered small dots of hyperfluorescence in the posterior pole and periphery during the angiography process with fluorescein leakage gradually with the prolongation of the angiography time.

Proposed Model Output (BLEU Score: 0.170): ”No significant abnormal fluorescence was seen in all phases of fluorescein fundus angiography angiography in the left eye.

Case 2:

Ground Truth: ”Left eye FFA: Fluorescence staining of the optic disc with indistinct borders. Retinal arteries are thin, veins are tortuous and dilated, some of the walls are fluorescently stained, capillary dilatation and increased permeability of small veins from the posterior pole to the mid-periphery and small scattered patches of capillary occlusions are seen in the mid-periphery. Capillary dilatation and increased permeability of the macula and cystic edema in the macula in the late stage.

Baseline Model Output (BLEU Score: 0.721): ”FFA of the left eye: The fluorescence image of the optic disc is approximately normal. The retinal arteries are thin, the veins are tortuous and dilated, some of the walls are fluorescently stained, the capillaries of the small retinal veins are dilated with increased permeability, a large number of laser spots can be seen from the mid-periphery to the periphery, and several vascular occlusion areas can be seen in the mid-periphery. Advanced edema in the macula.

Proposed Model Output (BLEU Score: 0.185): ”No significant abnormal dye leakage from macular choroidal vessels. No significant abnormalities in the choroidal vessels in the mid-periphery and periphery. No significant abnormal dye leakage from the optic disc and retinal vessels.

Case 3:

Ground Truth: ”FFA of the right eye: No abnormal fluorescence was seen in the optic disc and retinal vessels. Fluorescence was seen in the temporal punctate RPE pigment loss in the macula. No significant abnormal fluorescence was seen in the mid-peripheral retina.

Baseline Model Output (BLEU Score: 0.760): ”No significant abnormal fluorescence was seen in the macula. No significant abnormal fluorescence was seen in the mid-peripheral and peripheral parts of the retina. No significant abnormal fluorescence was seen in the optic disc.

Proposed Model Output(BLEU Score: 0.242): ”No significant abnormal fluorescence was seen in all phases of fluorescein fundus angiography angiography in the left eye.