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. Introduction

A. Background Information of FFA

Fundus Fluorescein Angiography (FFA) [1] is a diagnostic procedure used by ophthalmologists to examine blood circulation in the retina and choroid, which are layers of tissue in the back of the eye. The procedure consists of injecting a fluorescent dye, called fluorescein, into the patients bloodstream, usually through a vein in their arm. The dye travels through the circulatory system and reaches the blood vessels in the eye. A specialised camera equipped with filters to detect fluorescence emitted by the dye is used to take a series of photographs of the retina as the dye circulates through the blood vessels. Using these images radiologists can detect lesions, abnormalities and damages in the tissue, and then diagnose patients accordingly. This is a very time consuming process as outlined in [2] and radiologists could be spending this time in other areas such as attending to more patients.

The FFA-IR data set [3] 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.

B. Aim and Reasoning

The data set was created to improve medical report generation (MRG) which is the automatic generation of medical reports using a deep learning neural network [3]. The aim of this research is to improve the accuracy of MRG by implementing an object detection model to replace the feature extraction model in the MRG baseline code provided by [3].

Most MRG models are far too inaccurate in providing reliable results and this can be seen in the results section of [3]. By incorporating object detection methods into MRG models there is promise in enhancing the precision and localisation of abnormalities and the establishment of clearer links between identified objects and report descriptions. There are several compelling reasons for the research community to consider this integration.

Firstly, unlike MRG models that use a feature extractor and then focus on understanding the whole image, object detection prioritises on identifying the areas of importance within specific objects in the image. This nuanced difference potentially could lead to comprehensive image understanding and a more granular report generation based on specific identified elements.

Secondly, although incorporating an object detection module may significantly increase the complexity and computational resource requirements of the model, the promise of improved accuracy and precision in medical imaging may outweigh these challenges. Advances in computational power and resources, and their declining cost, make this more feasible now than ever before.

Furthermore, in the past expertly labelled bounding boxes for training an object detection model were rare. Fortunately, an increase in availability of high-quality, annotated medical imaging data sets, such as FFA-IR, have made this experiment more feasible.

While interpretability remains critical in medical applications, a more complex model incorporating object detection could offer superior explainability by clearly linking identified objects with descriptions in the report.

The potential advantages of merging object detection with MRG are immense, offering the promise of enhanced accuracy and practical value of the reports produced. By utilising the FFA-IR dataset, and the foundational MRG code provided in the report [3], which consists of a feature extractor and an encoder-decoder transformer, the suggested research aspires to embed an object detection model within the MRG code provided in [3]. This research will involve, choosing a suitable object detection model, incorporating it into the MRG code provided by [3], and selecting an appropriate method for evaluating the performance of the final medical reports.

This research holds substantial potential to propel advancements in the realm of MRG using medical images. Its focus remains fixed on boosting accuracy whilst also ensuring the explainability and reliability of the outcomes.

II. FFA-IR DATA SET

A. Overview: Why FFA-IR?

The research paper on FFA-IR [3] 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). 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.

III. MEDICAL REPORT GENERATION

A. Initial Approaches

The roots of MRG can be traced back to several early methodologies. In 2012, Varges et al. [4] 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 [5] 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 [6]. 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 [7] that proposed an encoder-decoder architecture that would later go on to become a standard for Medical Report Generation (MRG).

B. *Encoder-Decoder Architecture*

In the research report, Cho et al. [7], 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 [8].

The paper [9] 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 [8]. RNNs [10] 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 [9] the RNN used is called a Long Short-Term Memory (LSTM) [10], which is a very common choice of RNN [8].

LSTM [10] 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.

C. *Transformer encoder - decoder*

Transformers [11] 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 [11] uses an encoder-decoder framework, each composed of multiple layers of self-attention and feed-forward neural networks.

The encoder processes [11] 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 [11] 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 [12]. 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 [12].

D. Retrieval Approaches for MRG

Another approach is known as the Hybrid Retrieval-Generation Reinforced Model (HRGR) [13]. 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 [8], 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 [14] outperformed state of the art encoder-decoders.

E. 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 [15]. 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.

IV. Medical Report Generation Evaluation

A. 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 [16], CIDEr [17], METEOR [18], and ROUGE [19] aim to calculate the similarity between source and target sentences based on the occurrences of N-gram or word matching.

The FFA-IR dataset, which includes a range of medical reports, serves as a valuable resource for assessing the performance of these models. By comparing the generated reports with those in the FFA-IR dataset, the MRG models effectiveness can be evaluated.

B. NLG Evaluation Systems

1) BLEU: Bilingual Evaluation Understudy (BLEU) [16] measures the similarity between the candidate sentence and one or more reference sentences based on the n-gram overlap between them. BLEU calculates a score from 0 to 1, where higher scores indicate more similarity between the candidate and reference. However, BLEUs limitation lies in its reliance on the exact matching of n-grams, excluding potential synonyms or alternative phrasings that may also be correct.

2) CIDEr: Consensus-based Image Description Evaluation (CIDEr) [17] 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.

3) METEOR: Metric for Evaluation of Translation with Explicit Ordering (METEOR) [18] 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.

4) ROUGE: The Recall-Oriented Understudy for Gisting Evaluation (Rouge) [19] 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.

C. Summary

Based on a survey of Transformer-based MRG models [15], the majority use both CIDEr and BLEU. CIDEr and BLEU excel in scenarios where the precision and diversity in generated sentences are important [15]. 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.

V. Detection Algorithms

A. Overview of Object Detection

Object detection [20] 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 code provided by the FFA-IR article [3], it uses a CNN for feature extraction that then parses the extracted features to the transformer which performs MRG. The aim of this research is to replace this feature extractor CNN with an appropriate Object Detector CNN to attempt to improve the accuracy of the MRG.

Object Detector CNNs [20] 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.

B. 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 [21].

C. A brief history of Faster R-CNN: R-CNN and Fast R-CNN

1) R-CNN: R-CNN [22] (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.

2) Fast R-CNN: Fast R-CNN [23] 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.

3) Faster R-CNN: Faster R-CNN [24] 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.

D. Training Faster R-CNN

The creators of Faster-RCNN [24] 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.

1) 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.

2) 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 [24]. However, it does bring about a 25-50

3) 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.

E. One Stage Approaches

Unlike two-stage approaches that seperate object localisation and classification into distinct steps, one-stage approaches, such as SSD [25] and YOLO [26], 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 [21].

Single Shot MultiBox Detector (SSD) [25] 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) [26] 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.

It is widely accepted that two stage approaches are more accurate whereas one-stage approaches perform faster [21], [20].

VI. Detection Evaluation

A. Summary

In an literature survey on the performance metrics for object detection [21], 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.

B. Confusion Matrix

A confusion matrix [21] 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.

C. Precision and Recall

Two metrics are often calculated using the results from a confusion matrix and they are called precision and recall [21].

The precision is defined as:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{1}$$

Recall is defined as:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{2}$$

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 [21].

1) Precision-Recall Curve: To determine an appropriate balance between precision and recall a precision-recall curve [21] 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.

2) F1 Score: To determine the quality of the precision and recall the F1 score [21] 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.

D. 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

E. Method of Implementation

To implement these methods into the current pipeline provided by the FFA-IR article [3], 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 [27] where Detectron2 is used to evaluate the performance of object detection for autonomous vehicles.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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).

[5] 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.

[6] 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.

[7] 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.

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

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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).

[15] 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.

[16] 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).

[17] 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).

[18] 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).

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

[20] 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.

[21] 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.

[22] Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. InProceedings of the IEEE conference on computer vision and pattern recognition 2014 (pp. 580-587).

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

[24] 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.

[25] 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.

[26] 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).

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