Journal of Computer Science

Original Research Paper

Automated Medical Image Captioning with Soft Attention-Based LSTM Model Utilizing YOLOv4 Algorithm

Paspula Ravinder and Saravanan Srinivasan

Department of Computer Science and Engineering, School of Computing, Vel Tech Rangarajan
Dr. Sagunthala R&D Institute of Science and Technology, Chennai, Tamil Nadu, India

Article history
Received: 02-05-2023
Revised: 07-06-2023
Accepted: 04-09-2023

Corresponding Author:
Paspula Ravinder
Department of Computer Science and Engineering, School of Computing, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, Tamil Nadu, India
Email: vtd703@veltech.edu.in

Abstract: The medical image captioning field is one of the prominent fields nowadays. The interpretation and captioning of medical images can be a time-consuming and costly process, often requiring expert support. The growing volume of medical images makes it challenging for radiologists to handle their workload alone. However, addressing the issues of high cost and time can be achieved by automating the process of medical image captioning while assisting radiologists in improving the reliability and accuracy of the generated captions. It also provides an opportunity for new radiologists with less experience to benefit from automated support. Despite previous efforts in automating medical image captioning, there are still some unresolved issues, including generating overly detailed captions, difficulty in identifying abnormal regions in complex images, and low accuracy and reliability of some generated captions. To tackle these challenges, we suggest the new deep learning model specifically tailored for captioning medical images. Our model aims to extract features from images and generate meaningful sentences related to the identified defects with high accuracy. The approach we present utilizes a multi-model neural network that closely mimics the human visual system and automatically learns to describe the content of images. Our proposed method consists of two stages. In the first stage, known as the information extraction phase, we employ the YOLOv4 model to extract medical image features efficiently which is then transformed into a feature vector. This phase focuses primarily on visual recognition using deep neural network techniques. The generated features are then fed into the second stage of caption generation, where the model produces grammatically correct natural language sentences describing the extracted features. The caption generation stage incorporates two sub-models: An object detection and localization model, which extracts information about objects present in the image and their spatial relationships, and a sophisticated deep Recurrent Neural Network (RNN), which utilizes Long Short-Term Memory (LSTM) units, enhanced by an attention mechanism, to generate sentences. This attention mechanism enables each word of the description to be aligned with different objects in the input image during generation. We evaluated our proposed model, using the PEIR dataset. Various Performance metrics including Rouge-L, Meteor score, and Bleu score were evaluated. Among these metrics, the BLEU score obtained using this model was 81.78%, while the METEOR score achieved was 78.56%. These results indicate that our model surpasses established benchmark models in terms of caption generation for medical images. This model was implemented using the Python Platform, making effective use of its capabilities and PEIR dataset. We compared its performance with recent existing models, demonstrating its superiority. The high BLEU and METEOR scores obtained highlight the effectiveness of our suggested model excels in producing precise and contextually rich descriptions for medical images. In summary, the model

Science Publications

© 2024 Paspula Ravinder and Saravanan Srinivasan. This open-access article is distributed under a Creative Commons Attribution (CC-BY) 4.0 license.

performs exceptionally well in this regard. Overall, the development of this model provides a promising solution to automate medical image captioning, addressing the challenges faced by radiologists in managing their workload and improving the precision and dependability of generated descriptions.

Keywords: Automatic Medical Image Captioning, Deep Learning, Wiener Filtering, Color Channel, YOLOv4, Hyper-Parameter Tuning, Hybrid Attention, Long-Short-Term Memory

Introduction

Physicians have relied heavily on medical imaging as a resource for treatment and diagnosis (Xiong *et al*., 2019; Al Duhayyim *et al*., 2022). The method can be utilized in the medical profession to create health records from CT or X-ray images (Lee *et al*., 2022). Uses for image captioning include text-based image retrieval, relevant keyword assignment, human-robot interactions, and assistance for those who are blind or visually handicapped. Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), template-based, deep learning-based, and other techniques have been created for classifying images (Park *et al*., 2021; Morra *et al*., 2019).

Image captioning is a focal point between natural language processing and computer vision. The process of image captioning is used for different types of tasks. Despite many impressive advances, image captioning is far from being a solved task. It still is a challenge to satisfactorily bridge the semantic gap between image and caption and to produce diverse, creative, and human-like captions. Although such systems have been trained on libraries of mundane images that are obtained by mining from the web, one can easily think of beneficial use cases. For example, these image captioning language models are used to enable blind individuals to receive visual information about their surrounding environment.

Deep Learning was introduced in the early 20th century after Support Vector Machines (SVM), Artificial Neural Networks (ANN), and many other neural networks became popular. Deep learning is becoming a subset of Machine Learning (ML) which is considered a subset of Artificial Intelligence (AI) in turn during its inception period, deep learning didn't draw much attention due to scalability and several other influential factors such as demand for huge compute power (Ayesha *et al*., 2021).

Problem Statement

In the field of medical imaging, the medical images are read and interpreted by specialized medical professionals, and their findings regarding each body of area examined are communicated via written medical reports. The process of writing medical reports usually takes around 10 min per report per patient. In a day the doctors have to write medical reports that number in the 100s which can take a lot of their time.

The proposed study aims to develop an effective automated medical image captioning mechanism for reducing the workload of doctors and thereby saving time and expenses. The proposed medical image captioning attention model system incorporates five important stages namely image acquisition, preprocessing, feature extraction, tuning hyperparameters, and caption generation.

The Major Contributions of the Proposed Work are given in Detail Below

1. To introduce a novel deep learning model (Soft attention-based LSTM) for medical image captioning to reduce the workload of doctors and thereby save time and expenses
2. To pre-process the image by performing image de-noising using WF, image resizing, and color channel conversion
3. To extract the essential features from the pre-processed image using the YOLOv4 (Ayesha *et al*., 2021) model that can effectively reduce the unwanted over-fitting issue in the network model
4. To classify the medical captioning using a hybrid attention-driven bi-directional long short-term model (Soft attention based LSTM) that effectively learns the features and produces outstanding performance effectively
5. To optimize the network model using the Flamingo Search Optimization (FSO) algorithm to avoid unwanted errors in the classified outcome
6. To implement the proposed method in the PYTHON platform and performance measures like accuracy, recall, precision, F-measure, and RMSE and compared with existing techniques to prove the efficiency of the proposed model

Once these systems are trained on medical image datasets, then it could be beneficial to provide physicians with useful information and doctors able to do diagnosis procedures quickly. Also, these systems are very useful to bring visual intelligence through the image-sentence search process. The advanced recommendation and visual assistant systems will use automatic image captioning. Early techniques like template generation and slot filling (Kulkarni *et al*., 2013; Li *et al*., 2011; Farhadi *et al*., 2010) and caption retrieval (Ordonez *et al*., 2011; Gong *et al*., 2014; Hodosh *et al*., 2013; Sun *et al*., 2015) are used for automatic image caption generation. Deep neural networks employ better results when compared to these early techniques (Kiros *et al*., 2014a-b; Karpathy and Fei-Fei, 2015; Karpathy *et al*., 2014). The very common deep neural architecture used in many methods is the encoder-decoder (Sutskever *et al*., 2014; Cho *et al*., 2014; Kalchbrenner and Blunsom, 2013). In the encoder-decoder approach, the total image captioning process is divided into two parts, one is the feature-extracting phase called as encoding phase, and coming to the decoding phase is useful to generate the descriptions with respect to the encoded or extracted features. Deep neural techniques learn the features by using CNN (Ren *et al*., 2015; He *et al*., 2016) or a combination of object detectors (Girshick *et al*., 2014; Piwigo, 1999; Gkioxari *et al*., 2015) with CNNs.

In image captioning, we use the encoder-decoder architecture shown in Fig. 1 like how they are used in neural machine translation since we map visual features to a sequence of tokens, analogous to mapping an input sequence of words to an output sequence for translation. The tokens are the words of a sentence in an array before they are used in the process of generating word embeddings. Early deep attentive image captioning models employed Convolutional Neural Networks (CNN) as encoders and Long Short-Term Memory networks (LSTM) (Simonyan and Zisserman, 2014) as decoders (Schmidhuber and Hochreiter, 1997; Xu *et al*., 2015). In most of the work published recently and reviewed in this survey, bottom-up attention (Cho *et al*., 2015) is used for visual feature extraction. Image captioning using deep learning usually involves supervised learning. Although recently, (Anderson *et al*., 2018; and Laina *et al*., 2019) have recently shown that training a deep learning model for image captioning could lead to desirable results using unsupervised learning, the best results still come from the models trained with supervised learning. Therefore, it is not necessary to categorize the papers reviewed for this survey under supervised learning and unsupervised learning. In supervised learning, we always train the model with a dataset containing examples labeled with the ground truth of the output. Different types of encoders feed different types of information to the attention mechanisms and language models (decoders). After performing a review of the evolution of attention mechanisms in image captioning, we categorize the state-of-the-art literature based on the types of attention mechanisms used alongside different kinds of encoders such as CNNs or CNNs and object detectors (bottom-up attention encoders) or combination of these with Graph Convolutional Networks (GCN) **(**Feng *et al*., 2019; Bruna *et al*., 2013), coupled with various types of decoders, primarily LSTMs (Defferrard *et al*., 2016) and Transformer (Schmidhuber and Hochreiter, 1997). The transformer is an encoder-decoder-based model that consists of Multi-Head Attention (MHA) and scaled-dot attention (Self-Attention). Our goal in this survey is to review and reveal the best practices in employing attention mechanisms for image captioning in deep neural networks, especially among the state-of-the-art methods that achieve better performance in comparison with earlier types of attention mechanisms such as spatial soft and hard attention or semantic attention.

After completion of training a CNN on the image classification method on a provided dataset like image-net (Deng *et al*., 2009), the last soft-max feedforward layer is used for image classification. This mechanism provides us with a CNN backbone that can be used for feature extraction from unseen images.

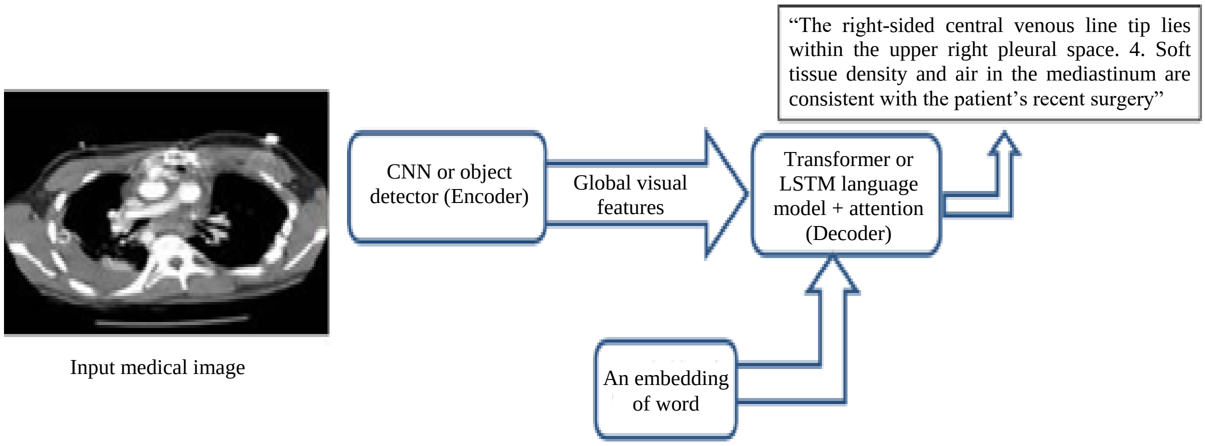

"The right-sided central venous line tip lies within the upper right pleural space. 4. Soft tissue density and air in the mediastinum are consistent with the patient's recent surgery"
CNN or object detector (Encoder)
Global visual features
Transformer or LSTM language model + attention (Decoder)
Input medical image
An embedding of word

Fig. 1: The task of image captioning using attentive deep learning models employing encoder-decoder architectures. The majority of deep learning models for image captioning use the encoder-decoder architecture

Table 1: Details of some previous methods

No	Reference	Method name	Advantages	Disadvantages	Hyperparameters
1	Gkioxari *et al*. (2015)	Image captioning model fused with object features	Effectively reduces the errors of an object category and improves the quality of generated sentences	The model cannot use for human action and of the processing time is high	Accuracy and recall rate
2	Al-Malla *et al*. (2022)	SPEA-II-based ATM	Encoder-decoder models have demonstrated notable improvements in efficiently extracting captioning from medical images	Suffers from hyperparameter tuning issues	
3	Papineni *et al*. (2002)	Attention-based Encoder Decoder deep architecture			The feature extraction method enhances the CIDEr score by 15.04%
4	Alabduljabbar *et al*. (2022)	The VaRaGlbResNet model- "An image a captioning system built on an end-to-end architecture, Utilizing an encoder-decoder framework enriched with attention mechanisms"	Our proposed model excels in of all performance metrics, demonstrating its effectiveness in enhancing caption quality by leveraging both spatial and global features, along with visual and textual attention		"Reported BLEU scores for four distinct N-gram lengths: N = 1, N = 2, N = 3 and N = 4
5	Vaswani *et al*. (2017)	The deep bidirectional SLTM model of three key components: A Convolutional Neural Networks (CNN) for encoding sentences and Multimodal LSTM (M-LSTM) for merging visual and textual representations into the results of the shared semantic space and generating sentences	Various methods have been proposed to enhance the depth of nonlinearity transition for the purpose of learning hierarchical visual language embeddings. These methods aim to increase the depth of nonlinearity transition in diverse ways	Needs more training time	Our Bi-S-LSTMA model demonstrates strong performance in various evaluation metrics. In the Flickr the 8K dataset, achieves a score of 19.4 for METEOR and 49.6 for CIDEr, surpassing Deep VSV model, which scored 16.7 for METEOR and 31.8 for CIDEr. Similarly, on the Flickr 30K dataset, our Bi-S-LSTMA model attains a METEOR score of 16.2 and a CIDEr score of 28.2, outperforming Deep VSV, which scored 15.3 for METEOR and 24.7 for CIDEr Furthermore, in the MSCOCO dataset, our Bi-S-LSTMA model excels with a METEOR score of 20.8 and a CIDEr score of 66.6, surpassing the Deep VSV model's scores of 19.5 for METEOR and 66.0 for CIDEr

Related Work

Singh *et al*. (2022) encoder-decoder models have demonstrated remarkable advancements in the efficient extraction of captions from medical images. The process begins by initially extracting image features through CNN layers additionally, the model leverages these extracted features to capture shape-related details and this iterative process is repeated to obtain end-level tokens. This study introduces an advanced architecture for deep learning, employing attention mechanisms within an encoder-decoder framework. This architecture harnesses convolutional features derived from a CNN trained on ImageNet (specifically, the xception model) and combines them with object features extracted from the YOLOv4 model, which has been pre-trained on the MSCOCO dataset (Al-Malla *et al*., 2022).

Alabduljabbar *et al*. (2022) proposed a comprehensive image captioning system developed using an encoder-decoder framework equipped with an attention mechanism. mechanism in the realm of image captioning systems, adopting an end-to-end methodology entails employing both the encoder and decoder components in a closely integrated fashion this system applies two attention mechanisms, the first to the visual features to concentrate on the image salient region and the second to the textual features to generate captions with more detailed information.

A revolutionary show attends and tells (ATM) paradigm has been created by Singh *et al.* (2022) and put into practice. There has been developed an encoder-decoder structure-based visual attention method. Issues with hyper-parameter adjustment plague SPEAII-based ATMs. To optimize the initial characteristics of an ATM system based on SPEAA-II was employed. Multiple experiments have demonstrated that ATM systems built on the SPEAK-II approach outperform existing models for captioning medical images. Only SPEA-II is utilized to fine-tune the model parameters. However, the ATM model is unable to produce an effective result due to the lack of an effective metaheuristic method. Table 1 contemplates the details of state-of-the-art methods.

Materials and Methods

In this study, we have formulated a technique capable of generating very relevant captions for the medical images we provided with the help of a given medical image data set.

Materials

Medical image dataset: Gather a dataset of medical images that you want to create captions for. These images should cover a variety of medical conditions and contain objects of interest that you want to detect and describe. The Pathology Education Information Resource (PEIR) digital library stands as a notable example among the publicly available medical image repositories, serving as a valuable resource for medical education. In the proposed model, the images are collected together with their portrayals in Gross sub-collection, ensuing in the PEIR Gross dataset. This dataset comprises 7,442 pairs of image captions from twenty-one dissimilar sub-categories of PEIR albums. In PEIR Gross, each caption consists of just one sentence, which sets it apart from the IU X- Ray dataset in terms of diversity. Moreover, the images have been annotated with tags from caption words with a maximum TF-IDF score.

Methods

The proposed model consists of two main parts: Encoding and decoding. The input to our model is a single image *I*, while the output is a descriptive sentence *S* consisting of *K* encoded words $S = \{1, w2,\dots, wK\}$. In the encoding part, firstly, we present a model that recognizes objects in the input image followed by a deep CNN to extract their locations, which reflect the spatial relationship associated. All the information will be represented as a set of feature vectors referred to as annotation vectors. The encoding part produces *L* annotation vectors, each of which is a D-dimensional representation corresponding to an object and its spatial location in the input image: $A = \{1, A2, \dots, AL\}$, $Ai \in R$ *D.3*. In the decoding part, all these annotation vectors are fed into a deep recurrent neural network model to generate a description sentence.

Object Detection with YOLOv4

Fine-tuning: Fine-tune the pre-trained YOLOv4 model on your medical image dataset. This involves updating the weights of the model using your dataset, so it becomes adept at detecting medical objects.

Inference: Use the fine-tuned YOLOv4 model to perform object detection on new medical images. This will identify and localize objects of interest in the images.

Image Captioning with LSTM

Preprocessing: Preprocess the detected objects (cropped regions of interest) and feed them as input to the LSTM captioning model. Additionally, you can use the original medical images as context for generating more informative captions.

Training: Train the LSTM model on your annotated dataset of image-object pairs and their corresponding captions. The LSTM learns to generate descriptive captions based on the input objects and their context.

Inference: Use the trained LSTM model to generate captions for the detected objects in new medical images.

Integration

Combine the object detection results from YOLOv4 with the generated captions from the medical image captioning model. Associate the captions with the detected objects and create a final report or visualization that includes both the object labels and their corresponding descriptions.

Evaluation

Evaluate the performance of your YOLOv4-LSTM model using metrics like BLEU score, METEOR, ROUGE, or human evaluation to assess the quality of the generated captions in comparison to ground truth captions.

Workflow of the Proposed Approach

In the Fig. 2 mentioned above shows outline of a possible workflow which includes two step processes.

Pre-Processing Stage

At the outset, an initial pre-processing phase is conducted to optimize the performance of medical captioning within the network model. In pre-processing, three major operations are performed namely image de-noising, rescaling, and color channel conversion. The noise from the raw dataset is completely removed using the WF technique followed by re-scaling and conversion operations are undertaken. Every image is resized to 256-256 pixels, and then 225/225 pixels are randomly cropped from it. The cropped images are then divided into a 3×3 grid of 9 patches, each one of size 75×75 pixels; each patch is then further cropped to size 64×64 pixels. Since our dataset had relatively bigger images than, Image-Net we tried to increase the image size to 384×384 pixels, and then randomly crop a patch of 339×339 pixels. The cropped image patch is then further converted into patches each of the image sizes is 113×113 of pixels, and then the image patches are further cropped to a final image size of 96×96 pixels.

Applying the Wiener Filter to Remove Noise, Using Python

Noise Removal Using IWF Technique

The images present in the raw dataset contain high noises and eliminating noises from the image is highly necessary to enhance the accuracy performance. Traditional Weiner filtering (Wang *et al*., 2016) technique does not eliminate noise from the background region and the produced outcome remains blurred. To overcome this issue, an Improved Weiner Filtering (IWF) technique is proposed in this study. The outcome of the noised image can be mathematically formulated as:

$$g_{(a,b)} = f_{(a,b)} \times u_{(a,b)} + n_{(a,b)} \tag{1}$$

The de-noised final outcome image $h_{(a,b)}$ from conventional *WF* can be mathematically interpreted as:

$$h_{(a,b)} = WF\left(g_{(a,b)}\right) \tag{2}$$

Here, implies the obtained image, denotes the degradation function, and indicates the noise (Gaussian). The de-noised final outcome image from conventional can be mathematically interpreted as:

$$\mu = \frac{1}{XY}\sum_{x,y\in\beta} p(x,y) \tag{3}$$

Thus, the noise gets removed without altering the originality of the image. The WF technique consists of both variance and mean pixel values in the size $x \times y$ mask matrix and it can be mathematically formulated as:

$$\sigma^2 = \frac{1}{XY}\sum_{x,y\in\beta} p(x,y)^2 - \mu^2 \tag{4}$$

Here, μ denotes the mean, σ^2 manipulates the variance of Gaussian noise in the image, x, y represents the adjacent area β in the mask, and $p(x, y)$ indicates the pixels in area β. The new pixels obtained from the *WF* technique can be mathematically formulated as:

$$R_{(x,y)} = \mu + \frac{\sigma^2 - \vartheta^2}{\sigma^2} * \left(p(x,y) - \mu\right) \tag{5}$$

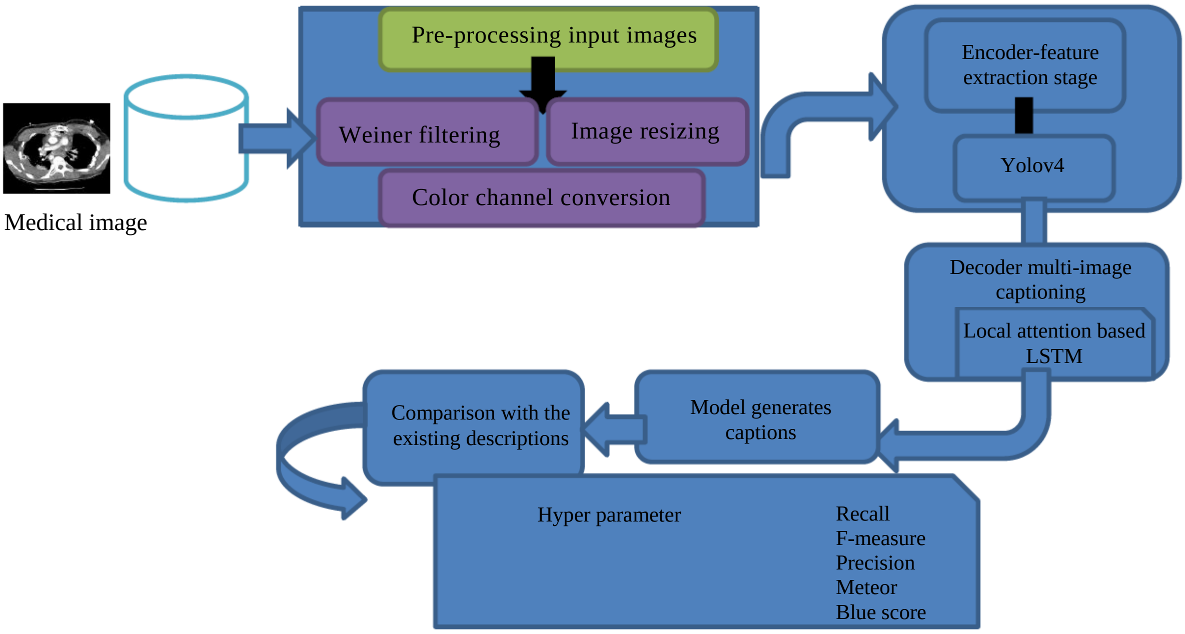

Medical image
Pre-processing input images
Weiner filtering
Image resizing
Color channel conversion
Encoder-feature extraction stage
Yolov4
Decoder multi-image captioning
Local attention based LSTM
Model generates captions
Comparison with the existing descriptions
Hyper parameter
Recall
F-measure
Precision
Meteor
Blue score

Fig. 2: Workflow of the proposed approach

Colour Channel Conversion

For evaluating the color channels, the order of pixels for each input image is determined. The proposed study developed the color vector for color channel evaluation, where the HSV color histogram is contemplated. The HSV color space is highly used in the color vision field. The Hue Saturation Value (HSV) value of each of the pixels in the given input image is transformed into RGB interpretation based on the following derivations:

$$H = \cos^{-1}\frac{\frac{1}{2}\left[(R-G)+(R-B)\right]}{\sqrt{(R-G)^2+\left((R-B)(G-B)\right)}} \tag{6}$$

$$S = 1-\frac{3\left[\min(R,G,B)\right]}{R+G+B} \tag{7}$$

$$V = \left(\frac{R+G+B}{3}\right) \tag{8}$$

Encoding Part

The proposed model's core insight is that when human beings try to describe an image using a sentence (combination of words), it's natural to first find objects and their relationships in the desired image. To imitate human beings, our encoding part has two steps, first, we use an object detection model is employed to identify objects within an image, and subsequently, a deep Convolutional Neural Network is utilized to determine their precise spatial positions.

Object detection: In the past few years, significant progress has been made in object detection. These advances are driven by the success of region proposal methods Hodosh *et al*. (2013) and Region-based Convolutional Neural Networks (RCNN) (Alabduljabbar *et al*., 2022). In our model, we choose faster RCNN (Deng *et al*., 2009) as an object detection model due to its efficiency and effectiveness in object detection tasks. Faster R-CNN is composed of two modules. The first module is a deep fully convolutional network that proposes regions, the second module is the fast R-CNN detector (Alabduljabbar *et al*., 2022) which uses the proposed regions. To generate 4 region proposals, the authors (Deng *et al*., 2009) slide a small network over the convolutional feature map output by the last shared convolutional layer. For each sliding window of the input convolutional feature map, this small network maps it to a lower dimensional feature. More explicitly, for every input image we detect *n* objects in an image, and each object is represented as a d-dimension vector:

$$\{j1, obj2, \ldots, objn\}, obji \in dd$$

Object localization: This part is designed to extract the information of objects' spatial locations which in turn reflect their spatial relationships. Sun *et al*. (2015) also considered the locations of different localized regions. However, they just added the boxes centrals with *x* location, *y* location, width, height, and area ratio with respect to the entire image's geometry to the end of the vector of each localized region. In this study, the implementation of extracting information from each object location is completely different from (Sun *et al*., 2015).

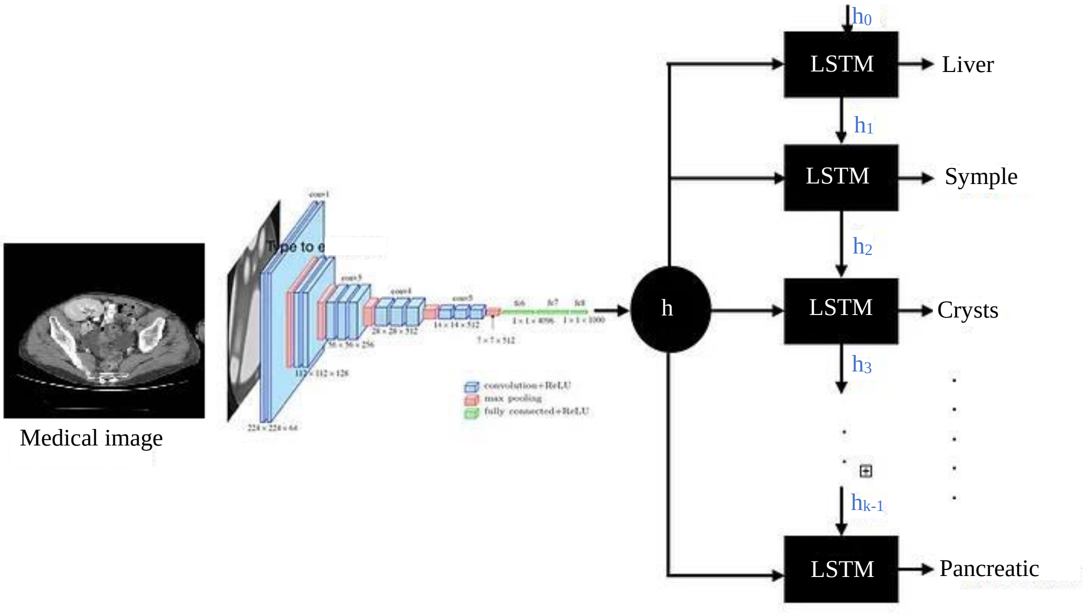


Fig. 3: A classic image captioning model

A classical medical image captioning model mentioned in Fig. 3 is a type of Artificial Intelligence (AI) system that combines computer vision and Natural Language Processing (NLP) techniques to generate textual descriptions or captions for medical images, such as X-rays, MRIs, CT scans, or histopathology images. Furthermore, we get another *n* vector of the *t* dimension in which each vector represents the information of the spatial location of each object:

$$\{c1, loc2, ..., locn\}, loci \in dt$$

Each annotation vector *Ai* consists of two parts: First, vector *obj* represents the feature of the object which particularly describes the contents of the image. Second, vector *loci* represent the feature of object location which tells us about the location of the individual object:

$$Ai = [ob; loci], Ai \in d^D, D = d + t \tag{9}$$

Training

YOLO (Ayesha *et al*., 2021) is an abbreviation for the statement 'YOLO: You Only Look Once' and is a real-time image processing system designed to detect and identify various objects within images. To achieve real-time object recognition, YOLO employs Convolutional Neural Networks (CNNs). "To perform detection, this method necessitates just a single forward pass through a neural network, as suggested by its name". Yolo has several variations, including YOLOv1, YOLOv2, YOLOv3, and YOLOv4 (Ayesha *et al*., 2021). YOLOv3 has the benefits of detecting speed and precision, as well as meeting the requirements of many applications. On the other hand, the Yolov3 model has difficulty with little items that appear in groups. YOLOv4 is a more advanced version of Yolov3 with the benefits of employing Dense-Net including boosting backpropagation, reducing gradient vanishing issues, enhancing learning, and removing the computational bottleneck. The neck component comprises the Spatial Pyramid Pooling (SPP) layer and the PANet Path Aggregation. Although it is fast and accurate enough to complete the job, it is still quite inaccurate when compared to other detection methods. It outperforms all other Yolo family members, including existing models such as the F-RCNN and AP-50. The YOLOv4 (Khan *et al*., 2022) is much smaller and faster than the YOLOv4 (Khan *et al*., 2022) and also generates more accuracy as well. The focusing layer in YOLOv4 minimizes the number of parameters, enhances the forward and backward speed, FLOPS, and CUDA memory, and reduces the influence of mean average precision. Feature extraction is one of the essential stages for extracting important features from medical images that is mainly responsible for enhancing the accuracy performance. Several existing techniques have been imposed to extract valuable features from the images. But those techniques face high time consumption and over-fitting issues. To overcome these problems, You Only Look Once (YOLO) version 4 model is proposed. The YOLOv4 (Khan *et al*., 2022) is one of the efficient object detectors that comes from the existing versions like YOLOv1, YOLOv2, YOLOv3, and YOLOv4 models (Ayesha *et al*., 2021). The proposed YOLOv4 model helps to overcome the regression problem by considering the overall objects into a single-level object effectively based on previous versions. However, this method is entirely different from the previous versions as it uses Pytorch instead of using darknet. The main backbone of the YOLOv4 (Khan *et al*., 2022) is the CSPDarkNet53. The suggested network architecture comprises three distinct layers: The head, neck, and backbone. The head layer is also known as the YOLO layer and the backbone layer plays a pivotal role in the extraction of features from the image. Due to YOLO (Ayesha *et al*., 2021) being a one-stage object detector, it performs both tasks simultaneously, which is also referred to as dense detection. In contrast, a two-stage detector handles these tasks independently and then combines the results, a method known as sparse detection.

Decoding Part

In this study, we describe a decoding part based on an LSTM network with an attention mechanism. The attention mechanism was first used in the neural machine translation area (Kiros *et al*., 2014a). Following the same mechanism, the authors (Kiros *et al*., 2014b; Karpathy and Fei-Fei, 2015; Ordonez *et al*., 2011) introduced it into the image processing domain whereas, (Ordonez *et al*., 2011) were the first to apply it to image captioning tasks. The key idea of the attention mechanism is that when a sentence is used to describe an image, not every word in the sentence is "translated" from the whole image but actually it just has a relation to a few sub-regions of an image. It can be viewed as a form of alignment from the words of the sentence to the sub-regions of the image. The feature vectors of these subregions are referred to as annotation vectors. Here in our implementation, sub-regions are referred to as the bounding box of objects, and annotation vectors are referred to as {*A*}, which is already discussed in the encoding part. In the decoding part, we follow (Ordonez *et al*., 2011) to use a Long Short-Term Memory (LSTM) network (Pavlopoulos *et al*., 2019) as a decoder. LSTM network products one word at every step *j* conditioned on a context vector, the previous hidden state *hj*-1, and the previously generated words *wj*-1 using the following formulations:

$$Inj = (WiEwj - 1 + Uihj - 1 + Zizj + bi) \tag{10}$$

$$fj = (WfEwj - 1 + Ufhj - 1 + Zfzj + bf) \tag{11}$$

$$cj = fjcj - 1 + \tanh\left(WcEwj - 1 + Uchj - 1 + Zczj + bc\right) \tag{12}$$

$$oj = \left(WoEwj - 1 + Uohj - 1 + Zozj + bo\right) \tag{13}$$

$$hj = oj\tanh\left(cj\right) \tag{14}$$

Here, *fj*, *cj*, *oj*, and *hj* represent the state of the input gate, forget gate, cell, output gate, and hidden layer respectively. W., U., Z., and *B*. have learned weight matrices and biases. *E* is an embedding matrix and *σ* is the logistic sigmoid activation. The context vector *zj* is derived from annotation vector *A*, where *i* ranges from 1 to *n*, representing the feature vectors of various objects:

$$zj = \sum aji\, ni = 1Ai \tag{15}$$

where, *αji* is a scalar weighting of annotation vector *Ai* at time step *j*, defined as follows:

$$eji = fa\left(Ai, hj - 1\right) \tag{16}$$

$$aji = \exp\left(eji\right)\sum \exp\left(ejk\right)nk = 1 \tag{17}$$

$$\sum aji\, ni = 1 = 1 \tag{18}$$

where, *fatt* is a multilayer perceptron conditioned on the previously hidden state *hj*-1. The positive weight *αji* can be viewed as the probability that the word generated at the time step "translated" from object *i*. We predict the next word *Wj* with a soft-max layer, the input of it is the context vector, the previously generated word, and the decoder state:

$$h : p\left(wj\right) \propto \exp\left(Lo\left(Ewj - 1 + Lhhj + Lzzj\right)\right) \tag{19}$$

where, *Lo*, *E*, *Lh*, and *Lz* are learned parameters.

Image Captioning Using LSTM

The obtained image captioning is effectively classified by employing the proposed hybrid attention-based Long Short-Term Memory (LSTM) technique, which is fed with the extracted features. Traditional LSTM technique (Simonyan and Zisserman, 2014) consumes high training time and is also often prone to high over-fitting issues. To overcome this issue, the hybrid attention mechanism is hybridized with the LSTM technique which can enhance the model performance effectively. The LSTM, short for Long Short-Term Memory, belongs to the category of Recurrent Neural Networks (RNNs) and plays a pivotal role in addressing the intricacies of sequential data.

Layers and Gate Operations

The LSTM model comprises three layers: The input layer, the hidden layer, and the output layer. The input layer plays a crucial role in providing input information in both left-to-right and right-to-left directions. The hidden layers in this network model are present in between input and output layers that help to perform complex nonlinear functions to process the data effectively. In addition, the LSTM consists of three gates namely input, output, and forget gate. The input gate decides the amount of information to be there in the current cell state. The forget gate plays a crucial role in managing data flow by filtering out unnecessary information from the previous state stored in the memory cell. Conversely, the output gate determines which information should be forwarded to the subsequent time step.

YOLOv4, or "You Only Look Once Fig. 4," is a real-time object detection system that builds upon the success of its predecessors, YOLOv1, YOLOv2 (YOLO9000), and YOLOv3. YOLOv4 is known for its impressive speed and accuracy in object detection tasks. It was created by Alexey Bochkovskiy, the author of the YOLOv4 repository on GitHub. In LSTM both forward and backward propagation operations are undertaken that can effectively minimize the over-fitting issues in the network model. Figure 5 determines the architecture of the Hybrid attention-based LSTM model.

The sequential data series can be expressed as, P_{task} = [p_0, p_1, p_2,…..p_t, …..p_n] here, p_t representing the *t* steps at p_{task}. Then the model evaluates the hidden cell outcome as H_{t-1} for each step and the final outcome is represented as O_{task} and each output O_t is determined using H_t. The forward propagation process can be mathematically formulated as:

$$a_t = \tanh\left(up_t\right) \tag{20}$$

$$H_t = wH_{t-1} + a_t + bias_H \tag{21}$$

$$O_t = vH_t + Bias_O \tag{22}$$

Here, *u* represents the input layer weight, *w* indicates the hidden layer weight, *v* indicates the output layer weight, and *u*, *w*, and *v* are given to each step respectively. The proposed A LSTM model effectively prevents the vanishing gradient problems and enhances each cell state operation effectively with the presence of an attention layer. Some of the cell state operations like forget gate *F*, input gate, *A*, and cell state C. The LSTM having each cell state operations can be mathematically interpreted as:

$$F_t = \left(w_{FH}H_{t-1} + w_{Fp}p_t + bias_F\right)\sigma \tag{23}$$

$$A_t = \left(w_{aH}H_{t-1} + w_{ap}p_t + bias_A\right)\sigma \tag{24}$$

$$C_t = F_tC_{t-1} + A_t\tanh\left(w_{CH}H_{t-1} + w_{Cp}p_t + bias_C\right)\sigma \tag{25}$$

$$Z_t = \left(w_{ZH}H_{t-1} + w_{Zp}p_t + bias_O\right)\sigma \tag{26}$$

$$H_t = Z_t\tanh\left(C\right) \tag{27}$$

The Sequence is as Follows

Input □ CSPDarknet53 □ (Backbone) □{SPP Block+ PANet} (Neck) □ YOLOv4 (Head)

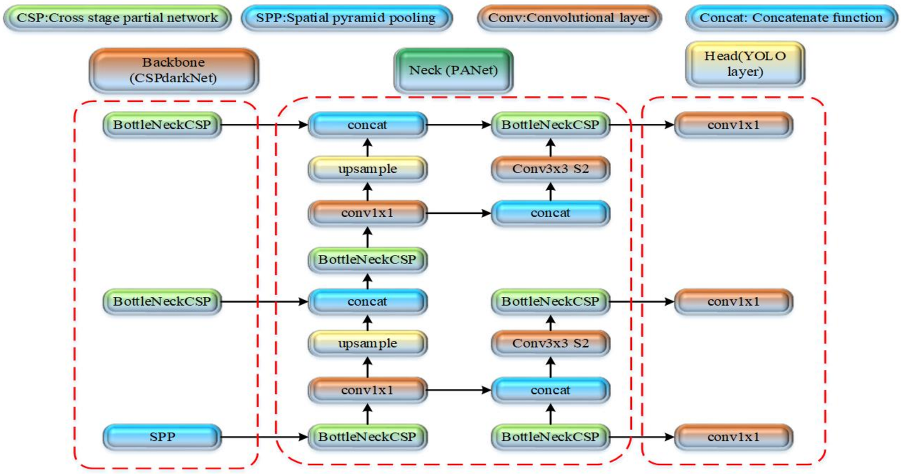


Fig. 4: YOLOv4 architecture

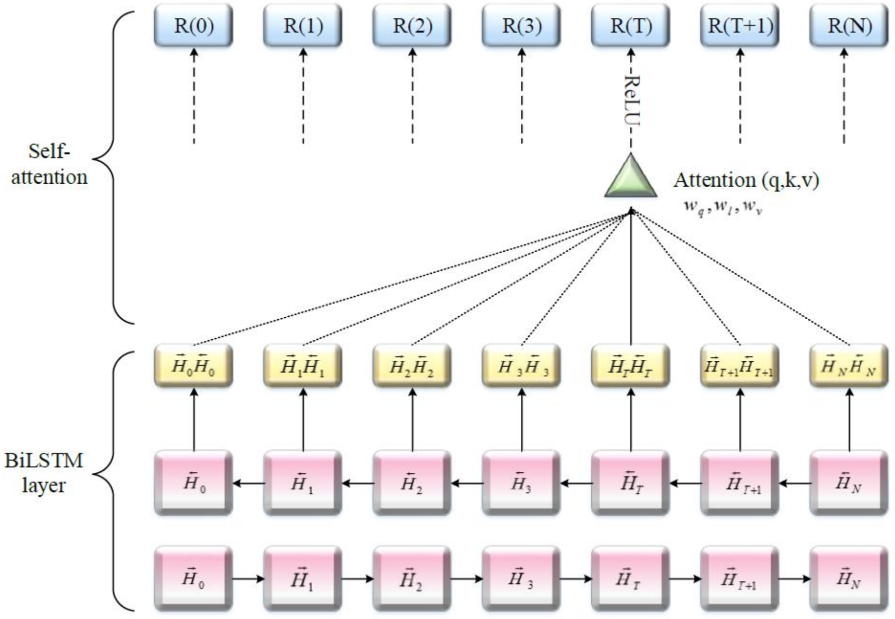


Fig. 5: Architecture of hybrid attention-based LSTM model

Beam Search

After completion of the training process of the image captioning model, the trained model is ready to provide the corresponding captions in the inference stage. Once feeding the <start> token to the language model-LSTM, the caption generation process will be continued until the model identifies the <end> token. The searching method beam search is selected to identify the next token in the inference stage, enabling the model to choose the sequence with the best overall score among several candidate sequences conditioned on the input image. At each decoding step, the search algorithm considers a set of top *k* sentences at time *t* as possible candidates to generate a sentence at time *t* +1 (*k* is the beam size). This process is repeated until the sequence with the highest overall score is generated.

The encoding part first extracts the information about objects (left) and their spatial relationships (right) in the image, and then the decoding part generates words based on these features.

Results and Discussion

The aim of image captioning is to generate a coherent sentence that conveys the visual content of an individual image. In this study, we utilize the prominent PEIR dataset (Bochkovskiy *et al*., 2020) to showcase the effectiveness of the proposed approach.

Dataset and Metrics

The Pathology Education Information Resource (PEIR) digital library stands as a notable example among the publicly available medical image repositories, serving as a valuable resource for medical education Piwigo (1999). In the proposed model, the images are collected together with their portrayals in Gross sub-collection, resulting in the PEIR Gross dataset. This dataset comprises 7,442 pairs of image captions from twenty-one dissimilar sub-categories of PEIR albums. In PEIR Gross, each caption consists of just one sentence, which sets it apart from the IU X-ray dataset in terms of diversity. Moreover, the images have been annotated with tags from caption words with a maximum TF-IDF score.

Data Preparation

To generate captions using the suggested model, it underwent training on the PEIR dataset. The given input data in the form of images should be appropriately prepared. To get accurate results from the training and evaluation process, the training part of the dataset is pre-processed Separate the images and captions into a suitable format individually. This model performs preprocessing on all images, ensuring they conform to the appropriate format. It rescales pixel values to a range of [0, 1] and subsequently normalizes them using a specific method. _ = [0:485; 0:456; 0:406] and _ = [0:229; 0:224; 0:225] First, we calculate the average and standard deviation of the RGB channels for the images in the ImageNet dataset. Following this, we resize all the images to ensure uniformity. When employing faster R-CNN, the image features are extracted and then stored in a cache on the disk.

Evaluation Measurements

This study uses a set of metrics for evaluation that are very useful in the process of Image captioning and relies on the widespread utilization of BLEU metrics for automated text assessment. These metrics serve to gauge the alignment between a machine-generated caption, akin to an automatically produced description, and a human's portrayal of the same image. This evaluation method is often employed in image captioning to measure the level of correspondence between machine-generated and human-generated descriptions. The generated BLEU score insufficiently compensates for the deficiency in the recall. Experimental results of (Banerjee and Lavie, 2005) strongly support this claim. METEOR was specifically created to tackle the shortcomings pinpointed in BLEU. It assesses a translation by determining a score derived from direct word-to-word comparisons between the translated text and a reference translation (Banerjee and Lavie, 2005):

$$Fmean = \frac{10PR}{R+9P} \tag{28}$$

$$Penalty = 0.5 *(\frac{\#chunks}{\#unigrams\_matched})3 \tag{29}$$

As the quantity of chunks decreases to one, the penalty diminishes and its minimum limit is reached Determined by the count of matching unigrams, the parameters for this penalty function were established. Some experimentation with development data, but have not yet been trained to be optimal. Finally, the meteor score for the given alignment is computed as follows:

$$Score = Fmean1(* - Penalty) \tag{30}$$

Quantitative Results

This section provides quantitative findings to illustrate the effectiveness of the proposed model. We compare the suggested technique to "Seven cutting-edge models in a multi-comparative analysis." In the attention-based, the language model, which relies on LSTM architecture, initially incorporates provided image features that have been extracted from the fully connected layer. A YOLOv4 model. Soft-Attention mechanism it employs a YOLOv4 model to extract regional representations, specifically focusing on the final convolutional layer. Then, it leverages an attention-based LSTM language model to decode each word at every timestep, depending on the chosen representations (Khan *et al*., 2022). In the context of multiple-instance learning, MSM introduces inter-attribute correlations and explores various methods for incorporating detected characteristics and image representations into an LSTM-based language framework (Khan *et al*., 2022). In the case of attribute-driven processing, a YOLOv4 LSTM architecture is employed, coupled with a visual attention mechanism for attribute detection, to capture co-occurrence dependencies among attributes (Khan *et al*., 2022). For visually grounded image captioning, the NBT architecture is utilized, enabling the generation of free-form natural language descriptions while also providing clear localization of objects within the image (Khan *et al*., 2022). Additionally, in medical imaging, a GNN is employed to model the relationships between objects and regions within a given medical image. In visual context-aware soft-attention, which considers the visual relationship between regions of interest to enhance the portrayal of visual content within the image (Khan *et al*., 2022), we seek to capture the subtle intricacies of various visual associations that may be discerned creating connections among objects and effectively utilizing these visual relationships.

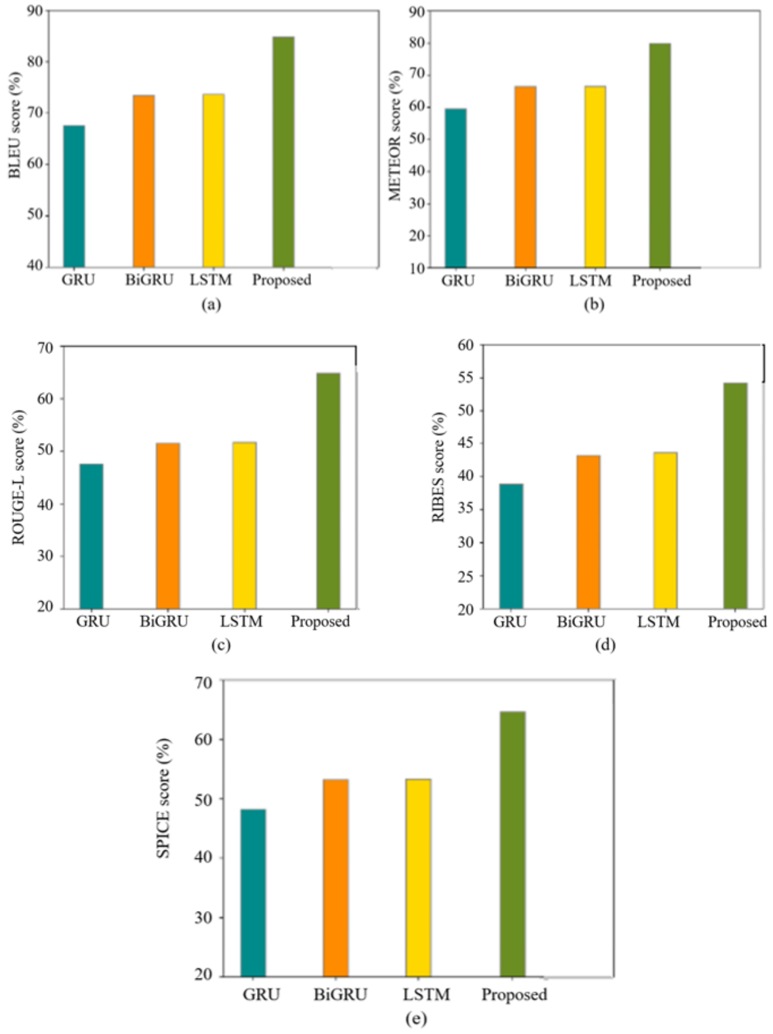
Qualitative Results

In this study, the proposed qualitative results of the model are quite impressive. The proposed model generates coherent and meaningful descriptions for a wide range of medical images.

When compared to YOLOv3 and other previous models, the proposed model MODEL with YOLOv4-ATT-LSTM has an improvement in the Bleu score, RIBESs score and Meteor score Table 2 exemplifies the evaluation outcomes of the proposed MODEL and existing models in terms of diverse metrics. In the table, the proposed MODEL has achieved better outcomes than the existing ones for medical image captioning. The proposed MODEL has the system capability to detect the captions based on the input samples. Despite the existing models, GRU has accomplished much fewer outcomes in the context of evaluating text quality and similarity, we consider metrics such as BLEU, METEOR, ROUGE-L, RIBES and SPICE score.

Figures 6 (a-e) describes the graphical representation of evaluated metrics in the proposed model.

The proposed model has reached the supreme BLEU score in captioning the image, which is perceived in the graphical illustration. The obtained METEOR score through a proposed MODEL is about 78.56% which is greater than the existing models such as GRU of 58.48%, Bi-GRU of 66.23%, LSTM of 66.36%, Alternatively, the performance evaluation of the SPICE score, ROUGE-L score and RIBES score with proposed MODEL and existing models are characterized in Fig. 6. In the graphical illustration, it is discovered that the proposed MODEL has obtained an extreme RIBES score of 55.36% in contrast with existing models owing to the accurate detection of image caption. The RIBES score obtained by existing models includes 38.85% for GRU, 36.56% for Bi-GRU, 43.23% for LSTM, and 43.62% respectively. Henceforth, the presented system has been shown to be versatile in generating image captions based on quantified medical input images.



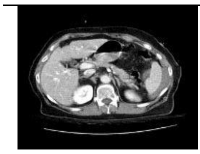
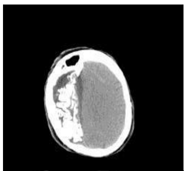
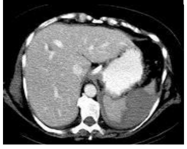
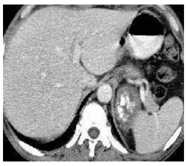
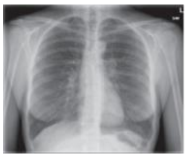
(a) (b) (c) (d) (e)

Fig. 6: Performance comparison of proposed and existing methods (a) BLEU; (b) METEOR; (c) ROUGE-L; (d) RIBES; (e) SPICE

Table 2: Proposed model performance evaluation by using BLEU score, METEOR score, ROUGE-L score, RIBES score and SPICE score

	Performance evaluation %				
Methods	BLUE	METEOR	ROUGE-L	RIBES	SPICE
GRU	67.54	58.48	47.56	36.56	49.56
Bi-GRU	76.58	66.23	52.56	43.23	54.26
LSTM	77.36	66.36	51.23	43.62	53.18
Proposed model	81.78	78.56	65.25	55.36	65.36

Table 3: Insists on the predicted and reference caption with respect to the input sample image

	**Reference Caption:** Filling defects in the segmental arteries of the right and left lower upper lobes consistent with pulmonary emboli **Predicted Caption:** Femoral herniation appendicitis inflammatory changes differ seen within the left inguinal region which out to comprise primary to appendicitis within a left femoral appendiceal hernia no inflammatory change is seen within the peritoneum fatty liver diverticulosis of the sigmoid colon no same abnormality differs identified
	**Reference Caption:** Splenic laceration with active contrast extravasation **Predicted Caption:** Splenic laceration with extinct demarcation eructation
	**Reference Caption:** Splenic infarct; intraperitoneal hemorrhage. A sizable wedge-shaped region in the spleen shows reduced attenuation, indicative of an infarction. Additionally, there is the presence of fluid around the spleen **Predicted Caption:** Splenic infarct intraperitoneal hemorrhage at that place lives a small wedge-shaped region of elevated density in the spleen, indicative of a splenic infarct, is observed There is also notable variance in the perisplenic fluid, suggesting potential changes in the future
	**Reference Caption:** Peritoneal leiomyomatosis there is a general expansion in size and an increase in the quantity of multiple peritoneal omental soft tissue nodules. The soft tissue densities between the right hepatic lobe and kidney appear to remain largely unchanged. Unusual soft tissue is present in the right pelvic region near the cervix of the uterus, accompanied by several small bowel loops **Predicted Caption:** Peritoneal leiomyomatosis "Overall expansion in both size and the quantity of individual peritoneal omental fat tissue." nodule gentle tissue paper concentration between the left hepatic lobe and kidney differ especially altered normal hardened tissue within the incorrect pelvis adjacent to the uterus cervix and respective large intestine uncoil the loop
	**Reference Caption:** The heart appears to be of regular size and the cardio mediastinal silhouette appears within normal limits. The lungs exhibit marked expansion, causing a flattening of the hemidiaphragms. No focal airspace opacities, pleural effusions, or pneumothoraces are detected. Multilevel degenerative alterations are observed in the thoracic spine **Predicted Caption:** Predicted: Narrowing of the right hand individual iliac nervure Crataegus laevigata thurner syndrome episodic swelling of the right under structure and high leg

Tables 2-3 exemplifies the evaluation outcomes of proposed MODEL and existing models in terms of diverse metrics. In the table, the proposed MODEL has achieved better outcomes than the existing ones for medical image captioning. The proposed MODEL has the system capability to detect the captions based on the input samples. Despite the existing models, GRU has accomplished much fewer outcomes in the context of evaluating text quality and similarity, we consider metrics such as BLEU, METEOR, ROUGE-L, RIBES and SPICE score.

Simulation Outcomes of the Proposed MODEL

In this section, we delve into the results obtained from simulating a proposed model system designed for the automatic captioning of medical images. The proposed MODEL acquired medical input images from the provided dataset and performed pre-processing since the raw input images encompass more noise; subsequently, it can worsen the system performance. Therefore, the noises in the input images are eliminated in the pre-processing stage to enhance the image quality. After effective pre-processing, preprocessed images are fed as the input of the YOLOv4 model. In YOLOv4, a bounding box is generated to localize the target and augment the detection performance. Finally, the captioning process is performed through the Hybrid attention-based LSTM.

The proposed model has reached the supreme BLEU score in captioning the image, which is perceived in the graphical illustration. The obtained METEOR score through a proposed MODEL is about78.56% which is greater than the existing models suchas GRU of 58.48%, Bi-GRU of 66.23%, LSTM of 66.36%, Alternatively, the performance evaluation of the SPICE score, ROUGE-L score, and RIBES score with proposed MODEL and existing models are characterized in the figure. In the graphical illustration, it is discovered that the proposed MODEL has obtained an extreme RIBES score of 55.36% in contrast with existing models owing to the accurate detection of image caption. The RIBES score obtainedby existing models includes 38.85% for GRU, 36.56% for Bi-GRU, 43.23% for LSTM, and 43.62% respectively. Henceforth, the presented MODEL system has been shown to be versatile in generating image captions based on quantified medical input images.

Conclusion

This study presents a multi-model Deep Neural Network that automatically learns and describes the content of images. This novel model first extracts the information of objects and their spatial locations in an image and then a deep Recurrent Neural Network (RNN) based on LSTM units with an attention mechanism generates a description sentence. Each word of the description is automatically aligned to different objects in the input image when it is generated. The proposed model is more optimized compared to other benchmark algorithms on the ground that its implementation is totally made on human visual systems. We hope that this study will serve as a reference guide for researchers to facilitate the design and implementation of image captioning. In the future, our proposed model can be used to help clinicians as a second option, to increase confidence in the diagnosis.

Compared to existing methods the proposed one benefited from overall object detection extracted features from you only look once YOLOv4 and proved Reorganizing the extracted object tags can enhance the impact. Further enhancements to the proposed model are possible through this approach. Concatenating both objects to detect features from YOLOv4 and features extracted from convolution features using CNN. In the future, this proposed method can also benefit by adding some more Textual descriptions containing rich semantic information about objects in an image, instead of mere object layouts, which are essential for enhancing image understanding and interpretation and leads to generating meaningful medical image captions. Furthermore, more well-known techniques can be able to pre-process extracted tiny medical image pre-processing object features prior to inputting them into the decoder, and employing more advanced natural language models such as Bidirectional Long Short-Term Memory (BiLSTM) networks and meshed-memory transformers, can also be considered applied.

Acknowledgment

I would like to express my heartfelt gratitude to my supervisor Dr. S. Saravanan for providing invaluable feedback and guidance throughout the course of this research. I am also deeply appreciative of the constructive comments provided by the four reviewers from the journal of computer science.

Furthermore, I extend my thanks to the department of computer science and engineering at the school of computing, as well as the University of Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology for their support and resources.

Special recognition goes out to my friends and colleagues who offered unwavering motivation and support during the development of this study. Their encouragement was instrumental in its success.

Funding Information

The research paper focusing on medical image captioning, integrating YOLOv4 and an attention-based approach, was fully self-funded. This signifies that my study was carried out without any external financial support or sponsorship. I personally allocated resources, collected the dataset, and funded the research, leading to the development of this innovative approach to medical image captioning.

Author’s Contributions

Paspula Ravinder: I have made a substantial contribution to the concept or design of the article or the acquisition, analysis, or interpretation of data for the article and drafted the article or revised it critically for important intellectual content and also approved the version to be published lastly agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Saravanan Srinivasan: Served as an advisor and critically reviewed the study proposal.

Ethics

Writing the ethics section of a research article is essential to ensure transparency and compliance with ethical standards and guidelines. This section typically provides information about the ethical considerations and approval processes related to this research.

Conflict of Interest

No conflicts of interest that may have influenced this research. Ethical approval: I didn’t receive ethical approval. Human or animal subjects: This research doesn’t involve human subjects, animals, or both. Participant confidentiality: No participant confidentiality and privacy were maintained. Risks and benefits: No potential risks and benefits associated with this research. Compliance with ethical guidelines: State that the research adhered to relevant ethical guidelines, regulations, and laws. Data access and availability: In this research study, I have utilized the publicly available PEIR dataset. Reporting of ethical violations: In this study, no efforts were made to document any ethical violations that may have arisen during the research process. Compliance Statement: This research study adhered to all relevant ethical principles and guidelines.

References

Al Duhayyim, M., Alazwari, S., Mengash, H. A., Marzouk, R., Alzahrani, J. S., Mahgoub, H., ... & Salama, A. S. (2022). Metaheuristics optimization with deep learning enabled automated image captioning system. *Applied Sciences*, *12*(15), 7724. https://doi.org/10.3390/app12157724

Alabduljabbar, G., Benhidour, H., & Kerrache, S. (2022). Image Captioning based on Feature Refinement and Reflective Decoding. *arXiv Preprint arXiv*: 2206.07986. https://doi.org/10.48550/arXiv.2206.07986

Al-Malla, M. A., Jafar, A., & Ghneim, N. (2022). Image captioning model using attention and object features to mimic human image understanding. *Journal of Big Data*, *9*(1), 1-16. https://doi.org/10.1186/s40537-022-00571-w

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6077-6086). https://doi.org/10.48550/arXiv.1707.07998

Ayesha, H., Iqbal, S., Tariq, M., Abrar, M., Sanaullah, M., Abbas, I.,... & Hussain, S. (2021). Automatic medical image interpretation: State of the art and future directions. *Pattern Recognition*, *114*, 107856. https://doi.org/10.1016/j.patcog.2021.107856

Banerjee, S., & Lavie, A. (2005, June). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization* (pp. 65-72). https://aclanthology.org/W05-0909.pdf

Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). YOLOv4: Optimal speed and accuracy of object detection. *arXiv Preprint arXiv:2004.10934*. https://doi.org/10.48550/arXiv.2004.10934

Bruna, J., Zaremba, W., Szlam, A., & LeCun, Y. (2013). Spectral networks and locally connected networks on graphs. *arXiv Preprint arXiv:1312.6203*. https://doi.org/10.48550/arXiv.1312

Cho, K., Courville, A., & Bengio, Y. (2015). Describing multimedia content using attention-based encoder-decoder networks. *IEEE Transactions on Multimedia*, *17*(11), 1875-1886. https://doi.org/10.1109/TMM.2015.2477044

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv Preprint arXiv*:*1406.1078*. https://doi.org/10.48550/arXiv.1406.1078

Defferrard, M., Bresson, X., & Vandergheynst, P. (2016). Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in Neural Information Processing Systems*, *29*. https://papers.nips.cc/paper_files/paper/2016/hash/04df4d434d481c5bb723be1b6df1ee65-Abstract.html

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248-255). IEEE. https://doi.org/10.1109/CVPR.2009.5206848

Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., & Forsyth, D. (2010). Every picture tells a story: Generating sentences from images. In *Computer Vision-ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11* (pp. 15-29). https://doi.org/10.1007/978-3-642-15561-1_2

Feng, Y., Ma, L., Liu, W., & Luo, J. (2019). Unsupervised image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4125-4134). https://doi.org/10.48550/arXiv.1811.10787

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 580-587). https://doi.org/10.48550/arXiv.1311.2

Gkioxari, G., Girshick, R., & Malik, J. (2015). Contextual action recognition with r* CNN. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1080-1088).

Gong, Y., Wang, L., Hodosh, M., Hockenmaier, J., & Lazebnik, S. (2014). Improving image-sentence embeddings using large weakly annotated photo collections. In *Computer Vision-ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13* (pp. 529-545). https://doi.org/10.1007/978-3-319-10593-2_35

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770-778). https://doi.org/10.1109/CVPR.2016.90

Hodosh, M., Young, P., & Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, *47*, 853-899. https://doi.org/10.1613/jair.3994

Piwigo. (1999). PEIR Digital Library. https://peir.path.uab.edu/library/index.php?/category/106

Kalchbrenner, N., & Blunsom, P. (2013, October). Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 1700-1709). https://www.aclweb.org/anthology/D13-1176

Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3128-3137). https://doi.org/10.48550/arXiv.1412.2306

Karpathy, A., Joulin, A., & Fei-Fei, L. F. (2014). Deep fragment embeddings for bidirectional image sentence mapping. *Advances in Neural Information Processing Systems*, *27*. https://doi.org/10.48550/arXiv.1406.5679

Khan, R., Islam, M. S., Kanwal, K., Iqbal, M., Hossain, M. I., & Ye, Z. (2022). A deep neural framework for image caption generation using gru-based attention mechanism. *arXiv Preprint arXiv:2203.01594*. https://doi.org/10.48550/arXiv.2203.01594

Kiros, R., Salakhutdinov, R., & Zemel, R. (2014a, June). Multimodal neural language models. In *International Conference on Machine Learning* (pp. 595-603). PMLR. http://proceedings.mlr.press/v32/kiros14.html

Kiros, R., Salakhutdinov, R., & Zemel, R. S. (2014b). Unifying visual-semantic embeddings with multimodal neural language models. *arXiv Preprint arXiv*: *1411.2539*. https://doi.org/10.48550/arXiv.1411.2539

Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y.,... & Berg, T. L. (2013). Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(12), 2891-2903. https://doi.org/10.1109/TPAMI.2012.162

Laina, I., Rupprecht, C., & Navab, N. (2019). Towards unsupervised image captioning with shared multimodal embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 7414-7424). https://doi.org/10.48550/arXiv.1908.09317

Lee, H., Cho, H., Park, J., Chae, J., & Kim, J. (2022). Cross encoder-decoder transformer with global-local visual extractor for medical image captioning. *Sensors*, *22*(4), 1429. https://doi.org/10.3390/s22041429

Li, S., Kulkarni, G., Berg, T., Berg, A., & Choi, Y. (2011, June). Composing simple image descriptions using web-scale n-grams. In *Proceedings of the 15th Conference on Computational Natural Language Learning* (pp. 220-228). https://aclanthology.org/W11-0326.pdf

Morra, L., Delsanto, S., & Correale, L. (2019). Artificial intelligence in medical imaging: From theory to clinical practice. CRC Press. ISBN: 10-9781032176468.

Ordonez, V., Kulkarni, G., & Berg, T. (2011). Im2text: Describing images using 1 million captioned photographs. *Advances in Neural Information Processing Systems*, *24*.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311-318). https://aclanthology.org/P02-1040.pdf

Park, H., Kim, K., Park, S., & Choi, J. (2021). Medical image captioning model to convey more details: Methodological comparison of feature difference generation. *IEEE Access*, *9*, 150560-150568. https://doi.org/10.1109/ACCESS.2021.3124564

Pavlopoulos, J., Kougia, V., & Androutsopoulos, I. (2019, June). A survey on biomedical image captioning. In *Proceedings of the Second Workshop on Shortcomings in Vision and Language* (pp. 26-36). https://doi.org/10.18653/v1/W19-1803

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, *28*. https://doi.org/10.48550/arXiv.1506.01497

Schmidhuber, J., & Hochreiter, S. (1997). Long short-term memory. *Neural Comput*, *9*(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv Preprint arXiv:1409.1556*. https://doi.org/10.48550/arXiv.1409.1556

Singh, A., Krishna Raguru, J., Prasad, G., Chauhan, S., Tiwari, P. K., Zaguia, A., & Ullah, M. A. (2022). Medical image captioning using optimized deep learning model. *Computational Intelligence and Neuroscience*, *2022*. https://doi.org/10.1155/2022/9638438

Sun, C., Gan, C., & Nevatia, R. (2015). Automatic concept discovery from parallel text and visual corpora. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2596-2604). https://doi.org/10.48550/arXiv.1509.07225

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems, 27*. https://dl.acm.org/doi/10.5555/2969033.2969173

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*. https://doi.org/10.48550/arXiv.1706.03762

Wang, C., Yang, H., Bartz, C., & Meinel, C. (2016, October). Image captioning with deep bidirectional LSTMs. In *Proceedings of the 24th ACM International Conference on Multimedia* (pp: 988-997). http://dx.doi.org/10.1145/2964284.2964299

Xiong, Y., Du, B., & Yan, P. (2019). Reinforced transformer for medical image captioning. In *Machine Learning in Medical Imaging: 10th International Workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 10* (pp. 673-680). Springer International Publishing. https://doi.org/10.1007/978-3-030-32692-0_77

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., ... & Bengio, Y. (2015, June). Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning* (pp. 2048-2057). PMLR. https://doi.org/10.48550/arXiv.1502.03044