RESEARCH ARTICLE

INTERNATIONAL RESEARCH JOURNAL OF MULTIDISCIPLINARY TECHNOVATION

Asian Research Association

DOI: 10.54392/irjmt25510

Improving Medical Image Captioning with a Context-Aware Knowledge Graph Transformer Framework

Aarti Sahitya [a], Shilpa Shinde [a, *]

[a] Department of Computer Engineering, Ramrao Adik Institute of Technology/D.Y Patil Deemed to be university, Nerul, 400706, India

* Corresponding Author Email: shilpa.shinde@rait.ac.in

DOI: https://doi.org/10.54392/irjmt25510

Received: 15-04-2025; Revised: 27-08-2025; Accepted: 14-09-2025; Published: 28-09-2025

Abstract: In this paper, we proposed a context-aware knowledge graph transformer framework for improving the caption of chest X-ray images. Normally the role of a radiologist is to interpret the chest X-ray or MRI image and write a detailed summary of finding patterns in a report. To generate an automatic detailed summary of the image the proposed framework is divided into three steps. The first step captures the visual feature of images using computer vision algorithms as Resnet 50 and Alexnet. The Second step uses the knowledge graph layer is employed for calculating the similarity between the tokens based on angel and token overlap to generate context-aware meaning of each token. The third step utilizes the transformer-based decoder to generate the detailed caption. The performance of the proposed model is compared against existing baselines including LSTM, CONV2D, and BI-LSTM architectures. The Proposed model outperforms baseline models by achieving higher evaluation scores in terms of evaluation metrics as 63% (BLEU-1), 61% (BLEU-4), 79% (RIBES), 85% (precision), 82% (recall), 82% (SPICE),and 79% (METEOR) demonstrating its effectiveness in medical text summarization.

Keywords: Medical Image Captioning, Knowledge Graph, Transformers, Cosine Similarity, Jaccard Similarity

1. Introduction

Recent development in deep learning based computer vision models which helps in reading the images properly and generating the feature maps by applying deep convolutional layers for understanding the deep concept of images. Work has been going on generating the summaries from the image as human reads the image, understand it and write description of that image. This is possible for the normal images like an image of a girl playing football so the human can write easily description of that image but it creates difficulty for an human to understand the medical imaging and generate a descriptive sentences. So to understand the medical imaging properly and generate a summary out of it is carried out by machines by applying advanced computer vision models along with natural language processing. By applying only the models like LSTM, BILSTM, convolution neural network etc for generating captions is not sufficient as these models generates only one line captions of the inputed image. The existing models are not able to encode the image properly and not able to generate the descriptive story of that image. Such multi modal applications requires an additional work to improve the efficiency of model to understand the image for better generation of description. So here in this paper additional layer has been added between computer vision model and natural language processing transformer model known as knowledge graph layer whose works is to act as an embedding vector layer which collects image feature matrix from hybrid model comprising of resnet50 and alexnet, generates a textual feature vector by calculating the similarity index between features alias nodes and forward this generated feature vector to transformer layer to generate captions with whole line description. The NLP and transformer is one of the era of AI got huge success after the year 2010 when the computational capabilities of neural network has increased. The problem with medical image captioning [1] is bias problem arises when there is a mismatch between the training phase (where the model working upon with ground truth data) and testing phase (where the model has to create its own predictions).So if the difference occur in training and phase this leads to bias problem. There was a another problem occurs in terms loss-evaluation mismatch problem arises when using different evaluation metrics for performance measurement. The third problem arises of vanishing gradient problem which occurred during training phase when model has back propogated through the layers to reduce the error so gradient is lost and training phase is halted. The fourth problem arises of exploding gradient problem which occurs when gradient becomes large during back propogation operation brings the unstability

in neural network training. The fifth problem arises as object hallucination where image captioning model mistakenly identifies that object which is not available in the image. To measure the model has suffered from object hallucination problem using the CHAIR (caption hallucination assessment with image relevance) metric has been used which can work with BLEU and SPICE scores also.

The main contribution of this paper is

1. Resnet50 and alexnet model is utilized for extracting feature from chest xray images.
2. The proposed knowledge graph layer is proposed for creating context aware knowledge embeddings vector for finding similarity between tokens set for particular image.
3. The transformer decoder model is implemented for generating the detailed caption of image with contextual meaning.

2. Literature Review

In this study, the paper has studied in domain of image captioning and knowledge graph engineering. Their review as follows

In paper [2], the author has proposed method involves a three-stage design and utilizes a Siamese network to map keywords to sentence descriptions where that method is suitable for skin image captioning, particularly for small datasets and informal sentences. The results can provide a platform for medical students to learn and understand the meaning of skin lesions. The text also mentions the use of data pre-processing, augmentation, and labeling, as well as the involvement of various model architectures where the model suggested in this study achieves the better results. The text mentions that there were 1100 records in dataset out of which 800 is used for training and 300 for testing. The training data consisted of 49 positive samples and 255 negative samples. In paper [3] by zhanyu wang, hainian han, lei wang etal. where author has proposed transformer based model to generate diagnostic report under three criteria as aligned visual of a image and textual features, accurate multi-label discrimination, and word importance for report summarization. This model also performs better than other sate of art methodology but author also has provided the key for future research as incorporation of knowledge graph in the transformer based model which has capabilities to generate the strong relation between the medical terms. The normal image captioning task is easy than image captioning from radiology images in aspect that matching the disease keyword in natural language corpus. So author in this paper proposed two stage model where in first stage using a graph convolution model to train a multi-label classification network and second stage the LSTM decoder learns to attend to different finding on the graph to generate sentence level report generation. The author has also given future research direction to incorporate more graph structures [4]. In this paper the author has proposed Tandemnet2 to facilitate the interaction between visual & semantic knowledge for visual information distillation on natural and medical images to generate the textual description from images without intervening the underlying text availability with images. He suggested the future work to more exploration on medical image diagnosis [5].To understand the image of type medical domain then there is requirement of deep learning approach is needed, so author of this paper has proposed a baseline method to surface the classification of Covid-19, Influenza & other types of diseases. The author has received a better results for area under than curve for covid-19 and influenza detection. He directed the future work to develop the deep learning models which can assist the frontline radiologist [6]. The address the challenge of vanishing gradient problem, the author has proposed the efficient deep ensemble (DCNet) and evolved DCNet (EDCnet) to provide the explanatory information of medical images. The DCnet & EDCnet utilizes the resnet152, densenet 201. The future plan suggested to work on to improve the quality of images in form of visibility, noise and poor registration [7]. The author of this paper had proposed a model comprising of three components, 1) Encoder-It consist of CNN and multi-label classification task, predicting common observations and other medical concepts from the given images. 2) Decoder-It consists of hierarchical LSTM which generates sentence and word level description based on topic vectors received from encoder. 3) Reward module-It generates a reward via a reinforcement learning to measure the quality of a report generated at the output [8]. The author of this paper has proposed a relation-paranet framework which consist of topic encoder to explicit semantic consistency between the medical terms. An adaptive generator is employed to switch between template template retrieval and sentence generation, to generate the textual description of images from IUxray and CX-CHR dataset. The author has given further direction for generating the paragraph description of medical images [9]. In this study the author has proposed a model named a transformer based model designed for medical report generation (TrMRG) pretrained on the imagenet dataset and language model to enhance the quality of medical report specific to the IU dataset. This model achieves a better BLEU score compared to other state of the art models [10]. The MDINAP transformer-ewp model is proposed by author of this paper to generate medical report generation from chest xray dataset by avoiding average pooling to get the feature difference vectors to draft the report [11].This work introduces the concept of a topic scene graph, proposing a novel approach to learning important relations within an image. The methodology used in this paper is distilling the visual attention model on recognizing the objects from an inputted image in two branches as first order attention and second order

attention to supervise the relation between features to generate the linguistic scene graph and correspondence captions for particular image [12]. The author has proposed a novel approach for summarizing the Chinese meetings as KG based meeting summarization framework combination with encoder-decoder model and graph based transformer to learn the role-based semantic features of meetings [13]. The author has used two algorithms such as summary kg to calculate the top most nodes based on importance for computing top k nodes and used QuerySum KB for generating summaries from highly important nodes [14]. The author has used opinion graph algorithm for abstractive summary which extracts by analysing the highest ranking sentences for summary generation [15]. The author has used GG-NNs with sequence encoder to extract multi comment summarization (MCS) from abstractive text. This model achieves a better performance in ROUGE evaluation metrics. As text rank a graph based ranking algorithm for summarizing the content based on static word embeddings on BBC news data set of Kaggle [16-17]. The author has used a lauren framework which calculates the importance of nodes based on weightage of one node linking to another node in KG by assigning a score in form hubs and authority where by calculating hubs and authority score using SALSA algorithm [18].The author has proposed a spatio-temporal entity summarization using fusion strategy which collects information about entities based on time and location using fusion algorithm to evaluate summaries of entities, where entities are stored in knowledge graphs [19].The author has presented a new model for eliciting the captions from image using semantic hierarchical method for image to text summarization [20]. A paper by anubhav jangra had presented a automatic multi-modal summarization survey is given the future directions in 1) better fusion of multi-modal information should be developed to capture the semantic overlap acroos modalities. 2)Incorporation of novel evaluation metrics should be incorporated to judge the quality of summary generated from different modalities as text,audio,image,video etc.3)More datasets should be available in domain of medical report summarization, tutorial summarization, simplification summarization, slogan generation etc to develop the multiple potential application in that domain also.4)In today's world there is still research going on generating the complementary and supplementary summarization for abstractive multi-modal summarization. The author has provided further direction for developing multi modal summarization system for explainable and controllable mms, multi-lingual, data-stream, query based mms [21]. In this paper the author proposed a framework which combines visual and textual features to detect fake news posts of faked it dataset by crawling 1 million images through Reddit. The author has also given future directions that to combine a metadata and comments of the posts to enhance the user's creditability. There is a need to a research on cross domain generalization model to execute the multi-modal summarization from topics, websites and languages [22]. In this paper the author proposed a A2Summ a novel algorithm to generate a multimodal summary combined of video, image and text to improve the contrastive losses to exploit the inter-sample and intrasample cross-modality information [23]. Here author has proposed a novel framework for image summarization as CFSum which consist of three submodules as pre-filter,word-level complement and phrase-level complement to eliminate the impact of useless images. This CFSum algorithm well works to generate the summary from image but cannot be transferred to dual stream large models for heavy dimensional dataset [24]. In this article the author has carried out a research on TIB dataset for abstractive summarization of long multimodal documents for automatic videoconference summarization and also given the future direction as to combine video and text together as input to predict the abstractive summarization [25]. The author li.etal has proposed a novel method as elastic deep multi-view autoencoder with diversity embedding for detecting the Laplacian noise and outliers present in the Multiview dataset comprises of text and images. The suggested model is different traditional auto encoders which is more sensitive to Laplacian noise and outliers and better handling the Gaussian noise. The model embedded the multi-view auto encoder and graph constraint for maintaining the distinctive features across neighbors for reducing the error between incorrect neighbor relationships for improving noise robustness, view diversity and structural consistency via graph constraints [26].The author of this paper proposes a hybrid machine learning and deep learning model to improve the diagnostic accuracy with reducing the computational complexity of histopathological images of disease colorectal cancer. The mobile net V2 and densenet121 with transfer learning are used to extract robust features from histopathology images. The address the class imbalance syntactic minority over-sampling (SMOTE) techniques is utilized along with chi-square test for selecting the most relevant features and reduce dimensionality. The proposed model is tested on EBHI-seg (extended Bioimaging histopathological image segmentation) and a multi-class dataset. This composite approach achieves high performance in accuracy, precision, recall & F1 Scores by detecting optimized colorectal cancer from histopathology images [27].The author of this paper systematically review the latest research on diffusion based image captioning models such as DDPM,CLIP-diffusion-LM, Bit diffusion, Unibrain, versalite diffusion, SCD-net, Diffusion-RSICC in domain of medical imaging,self-attention, conditional diffusion and complex world-image relationships. The author suggested text based evaluations metrics as BLEU, METEOR, ROUGE, CIDER (consensus-based image description evaluation), SPICE (semantic propositional image caption evaluation) for measuring the n-grams overlap between generated caption with

reference captions. He also suggested the similarity based metrics which measures the semantic similarity between the generated caption with reference captions such as BERT score, CLIP score and Sentence-BERT(S-BERT). The diversity based evaluation metrics is also provided in this paper which assess how diverse or novel the generated captions are. This is important in diffusion based models which tend to produce varied and repetitive captions. The metrics of this type such as Distinct-n(distinct-1, distinct-2) for lexical diversity, vocabulary size, Self-BLEU for caption diversity and Novelty- proportion of captions or n-grams not seen in training data [28].

The previous studies carried out by the authors such as [3, 4-7] and [10] is primarily used only one or two pipelines for generating the summary from medical images which lacks in work of integrating the more concrete graph structures and the model for improving the quality of image visualization by reducing the noise from image which in thus improves the quality of medical text summarization. In contrast our proposed work uniquely integrates the hybrid resnet50 and Alex net model to improve the quality of image visualization by removing the noise from edges, corners, and contours along with knowledge graph layer which calculates the similarity between sentences based on angel similarity and token overlap similarity and producing the contextual aware graph embeddings which in turn fed to transformer model for generating the multiline summary of a chest Xray image.

The below table presents the proposed framework, notable contribution with domain, key techniques and dataset utilized on medical multimodal summarization task.

Table 1. Compress view of Literature review summary

Reference No	Author & Year	Proposed Framework	Domain	Notable Techniques Used	Dataset	Contribution /Future Work
[2]	Lin.Y *et al*,2023	3-stage Siamese-based keyword-to-caption model	Skin Image Captioning	Siamese network, augmentation	~1100 samples from medical dataet	Small dataset captioning; platform for medical students
[3]	Wang *et al.,* 2022	Transformer with multi-label + keyword fusion	Radiology	Visual-text alignment, keyword importance	Chest X-ray	Incorporate Knowledge Graphs into Transformers
[5]	Zhang *Y et al*, 2020	TandemNet2	Medical + Natural	Visual-semantic distillation	IU X-ray dataset available on Kaggle platform	Future work proposed on Deeper medical diagnostic understanding of X-ray images
[6]	Peng *Y et al*, 2020	Baseline DL for COVID-19, Influenza detection	Medical Imaging	CNN, multi-class classifier	Covid Image dataset from kaggle	Future directed by researcher to be done on Radiologist-assistive DL tools
[7]	Singh *et al.,* 2022	EDCNet (Efficient Deep Ensemble)	Medical Imaging	ResNet152 + DenseNet201	Medical images	Work in future can be carried out on Visibility, noise, registration enhancement for medical images
[8]	Hou *D, et al*, 2021	CNN encoder + LSTM decoder + RL reward	Medical Report Generation	Hierarchical LSTM + Reinforcement Learning	IU X-ray dataset	Sentence and word-level generation with reward is the contribution
[9]	Wang F *et al,* 2022	Relation-ParaNet	Medical Captioning	Template switching, topic encoder	IU X-ray CX-CHR	Future work to be done on Paragraph-level generation of medical images which we proposed in this research paper
[10]	Mohsan *et al*, 2023	TrMRG Transformer	Radiology	Pretrained Transformer (ImageNet+Lang)	IU Dataset	Outperforms baselines; high BLEU for proposed work. No future work is given
[11]	Wang *et al.,* 2022	MDINAP Transformer-EWP	Radiology	Feature-difference modeling	Chest X-ray	The proposed work avoids pooling with better report accuracy

						in generating the report
[12]	Wang *et al.*, 2021	Topic Scene Graph	General Image	Attention distillation, 2-branch model	Dataset is not mentioned in the paper	Visual-linguistic feature linking in generating scene graph
[13]	Qi *et al.*, 2022	KG-based Summarizer for Meetings	NLP (Chinese)	KG + Transformer	CSCWD	Role-based semantic learning for summarizing the meeting in Chinese context
[16]	Barman *et al.,* 2021	Graph-based static embedding summarizer	News (BBC)	TextRank, GloVe	BBC Kaggle	Lightweight graph methods for news text summarizer with knowledge graph embeddings
[18]	Jalota *et al.,* 2021	LAUREN Framework	KG Summarization	SALSA (Hubs/Authorities)	Dataset is not mentioned in the paper	Node ranking with hub-authority score for summarization task
[19]	Yang *et al.,* 2020	Spatio-Temporal KG Summarizer	KG	Time-location fusion + Graph learning	KG entities dataset	Time-sensitive entity summarization
[20]	Yao *et al.,* 2019	Hierarchical Parsing (HIP)	Vision-Language	Tree-LSTM, instance-region hierarchy	MSCOCO	Visual structure-based captioning on MSCOCO dataset
[21]	Jangra *et al.,* 2023	Multi-modal Summarization Survey	Multimodal	Survey	Multiple datasets can be used	MMS directions guidelines has been given for future work described in detail for literature review section
[23]	He *et al*., 2023	A2Summ Framework	Multimodal	Dual contrastive loss, cross-modality	Video + Text	Inter/intra-sample contrastive learning for cross-modality summarization task
[24]	Xiao *et al.,* 2023	CFSum	Image Summarization	Word & phrase complement modules	The dataset description is not given in paper	Avoids heavy dual-stream transfer on image summarization task
[25]	Gigant *et al.,* 2023	TIB Dataset & Model	Video Conference	Multimodal summarization	TIB dataset	Joint video-text abstractive summarization
[26]	Li *et al.,* 2022	Elastic Multi-view Autoencoder	Text + Image	Graph constraints, diversity embedding	Multi-view data	Noise-robust, Laplacian outlier resilience on text with image summarization task
[28]	Wang Z *et al,* 2023	Survey on Diffusion Captioning Models	Medical Vision	DDPM, BitDiff, SCD-Net	Multiple datasets available for captioning task	Captures diversity, semantics, robustness for medical vision datasets

3. Methodology

The methodology is divided in to three phases. The first phase is feature map detector from xray images using hybrid resnet50 and alexnet mode1.The second phase is extracting similarity between features based on angel and token overlapping collected from previous phase known as knowledge graph phase and third phase is transformer layer which generates summary from the similarity matrix generated from knowledge graph layer as shown in figure 1 below and the algorithmic steps for performing this methodology is indicated in table 2.

3.1 First Layer (Alexnet & Resnet 50)

The below architecture in figure 2 depict the usage of resnet50 and Alex net model.

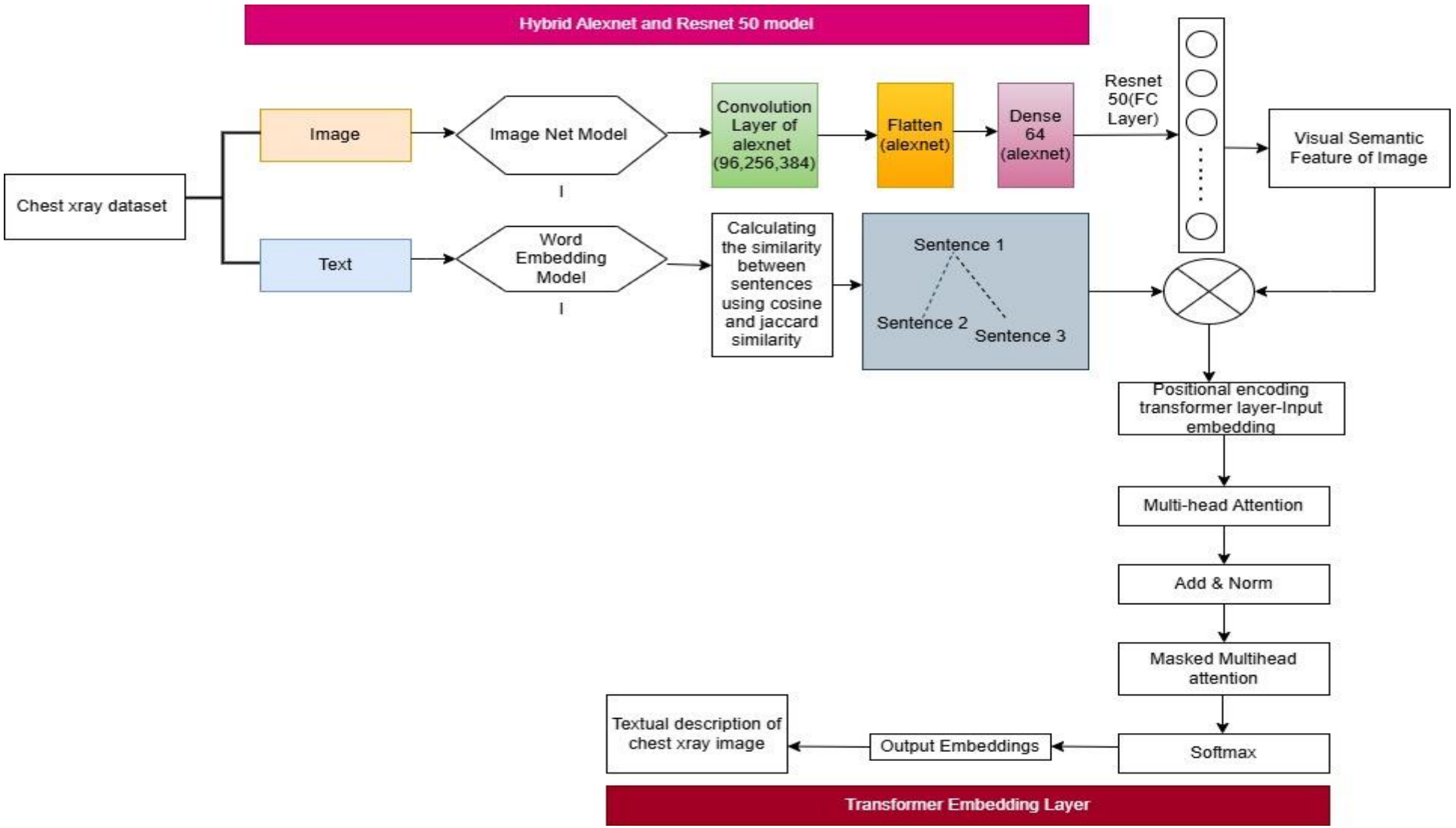


Figure 1. Overview of proposed work

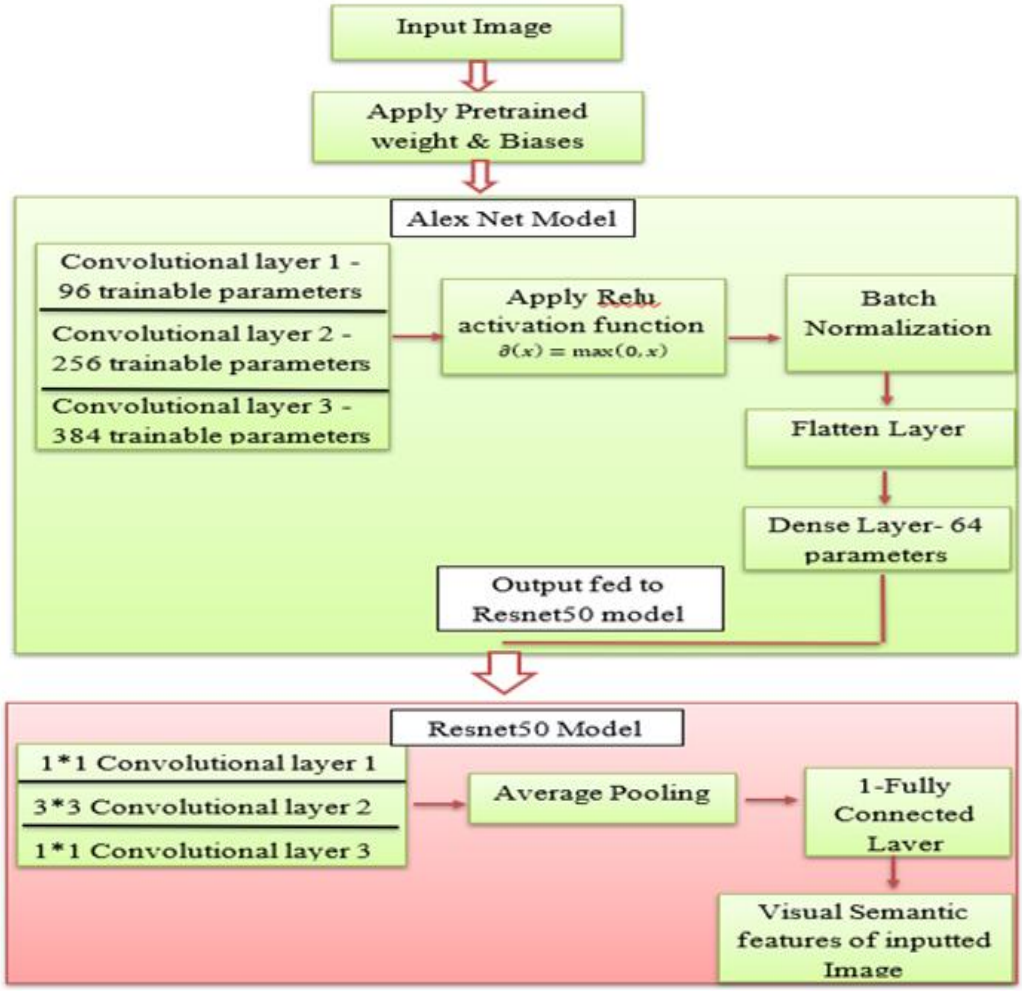


Figure 2. Schematic Workflow of Hybrid model (Alexnet+Resnet 50)

The motto behind the use of resnet50 and Alex net model, both models have their specialty in their own state of art, the resnet50 peculiarity lies by giving efficient accuracy with pretrained weights which gives the important features of an input image. Then we have utilized convolution 2D layers of Alex net model with flatten and dense layer it has peculiarity of producing 60 million parameters whereas resnet50 producing 26 million parameters means it refers to number of trainable weights and biases of convolutional and fully connected layer in each model architecture that influences both the capacity and complexity of a deep neural network i.e Alex net consists of 5 convolutional layers and 3 fully connected layers where each convolutional layer requires pretrained weights and biases of approx, 3.5 million and fully connected layer requires 4096*9216 trainable parameters results to approx, 37.7 million parameters, fully connected layer 2 requires 4096*4096 total trainable parameters results to approx 16.7 million parameters and fully connected layer 3 requires approx 4.1 million trainable parameters results by multiplying the 1000*4096 weights and biases. Resne50 26 million parameters results from trainable parameters requires for training the model as 1)initial convolutional (7*7) +Batch normalization(approx 9.4K trainable parameters),2)maxpool (0 parameters), 3)Residual block groupup1 convolutional (0.21M trainable parameters), 4)Residual block 2 convolutional layer(1.7M trainable parameters), 5)Residual block group 3 convolutional layer (approx 7.1M trainable parameters), 6)Residual block group 4 convolutional layer(14.8 M trainable parameers) and Fully connected layer(2.0M trainable parameters) which upon addition generates the 26M trainable weights and biases. To work on more parameters and to increase the accuracy combination of resnet50 and alexnet has been utilized. The model is using convolutional layer of 96,256,384 with flatten and dense layer along with by applying rectified linear unit (RELU) function for non-linearity mapping which has equation like

$$y = \partial(wx + b) \quad (1)$$

$$\partial(x) = \max(0, x) \quad (2)$$

The first formula is the basic adaptive parameters formula which states as y as output comprising of activation input as weights & bias. The second formula states that activation function as RELU. This model is using average pooling which gives results by taking mean of each patch. The schematic flow diagram of this hybrid model is observed in figure 2.

The diagram shown above in figure 2 explains the in-depth architecture of hybrid model consist of Alexnet and Resnet 50. Each model has its own peculiarity to deep learning an inputted image. We have utilized the 3 convolutional layers of Alexnet model, the reason for choosing the convolutional layers as its takes only 3.5 million trainable parameters out of 60 million parameters thus by not creating memory bottleneck while processing the images over google colab notebook or any python IDE where GPU limit is there. The converged images are then pass to the RELU activation function followed by batch normalization process. The normalized images are passed to the flatten layer and dense layer. We have not utilized the fully connected layer of alexnet because of huge parameter consideration. The generated output of Alexnet is passed to Resnet 50's convolutional layers, then to average pooling and then to fully connected layers which generates the visual semantic features embeddings of a particular image. The combination of Resnet50 and Alexnet is strategically designed to enhance image quality by reducing noise and better preserving essential features like edges, corners and contours which are crucial for accurate image interpretation in vision tasks. The Alexnet initial convolutional layers are good at filtering the noise from images in terms of edges, contours and corners thus by preserving the important low-level features. The strength of Resnet 50 fully connected layer consist of deep residual blocks which excels in eliminating the redundant information of an image and thereby focusing on producing meaningful visual semantic features of an image.

3.2 Second Layer (Knowledge graph layer)

In this layer the distance is calculated between the features collected from csv projections file of chest xray dataset as shown in figure 2. The distance between features is calculated using the sentence similarity method known as cosine similarity. The cosine similarity directly not works with the features in form of strings, that features is converted in to a vector which consist of frequency mapping of each features appeared. The cosine distance is calculated between the vectors using this formula (3) and (4). (Ai and Bi) given in the formula are the ith components of vectors A & B respectively. The similarity between features is calculated as the distance between the features measure in the angle of cos 90 as theta value. The greater the similarity between the features, the highest is the value of cosine similarity. This formula is derived by using the Euclidean dot product formula. The model uses jaccard similarity also to find the overlapping between tokens. The hybrid usage of cosine and jaccard improves the context aware of text. In the context of knowledge graph (KG) for medical text similarity calculating the angular similarity can prove to be effective for comparing similarity between two vectors present in sentence embeddings. The detailed view of this layer is shown in figure 3.

$$A.B = ||A||\,||B||\,cos\theta \quad (3)$$

$$Cos\theta = A.\frac{B}{||A||||B||} = \sum_{i=1}^{n} AiBi / \sqrt{\sum_{i=1}^{n} Ai^2}.\sqrt{\sum_{i=1}^{2} Bi^2} \quad (4)$$

Where

A.B = $||A||\,||B||\,cos\theta$ is the dot product of A and B

$||A||$=$\sum_{i=1}^{n} Ai^2$ is the magnitude of A

$||B||$=$\sqrt{\sum_{i=1}^{2} Bi^2}$ is the magnitude of B

Jaccard (A,B)=$\frac{|Atokens \cap Btokens|}{|Atokens \cup Btokens|}$ (5)

Where

$|Atokens \cap Btokens|$= Intersection of set of A tokens from Vector A with set of B tokens from vector B

$|Atokens \cup Btokens|$ = Union of set of A tokens from Vector A with set of B tokens from vector B

Hybrid(C,J)= $\lambda.cosine(A,B) + (1-\lambda).jaccard(A,B)$ (6)

Where

λ= is a weighting parameter that controls the contribution of cosine similarity and jaccard similarity in the final hybrid similarity score.

$\lambda.cosine(A,B)$ = is the dot product of cosine (A,B) with lambda

$(1-\lambda).jaccard(A,B)$ = is the dot product of jaccard (A,B) with the $(1-\lambda)$ is the weight assigned to vector of A and B

The importance of λ (lambda) is to create balance between two similarity factor as cosine and jaccard. The cosine similarity capture the semantic closeness based on word frequency and distribution which is dependent on vectors calculation and jaccard similarity captures the precise token overlap which is dependent on set based calculation. So here lambda is used as control parameter between cosine and jaccard. The detailed flowchart shown in figure 3 explains how knowledge graph layer will generate the graph embeddings of textual sentences in the dataset. The reason for integrating the cosine and jaccard similarity is the strength of cosine similarity is to find the angular distance between the two vectors thereby generating the semantic and distributional similarity between feature vector and the strong point of jaccard similarity is to measure the token overlap between two vectors which consist of overlapping entities and synonyms specially in medical text for eg the condition "opacity" and " pneumonia" is properly matched based on token overlap. The combination of two improves the coherence, context-awareness and factual accuracy of medical image multiline captioning.

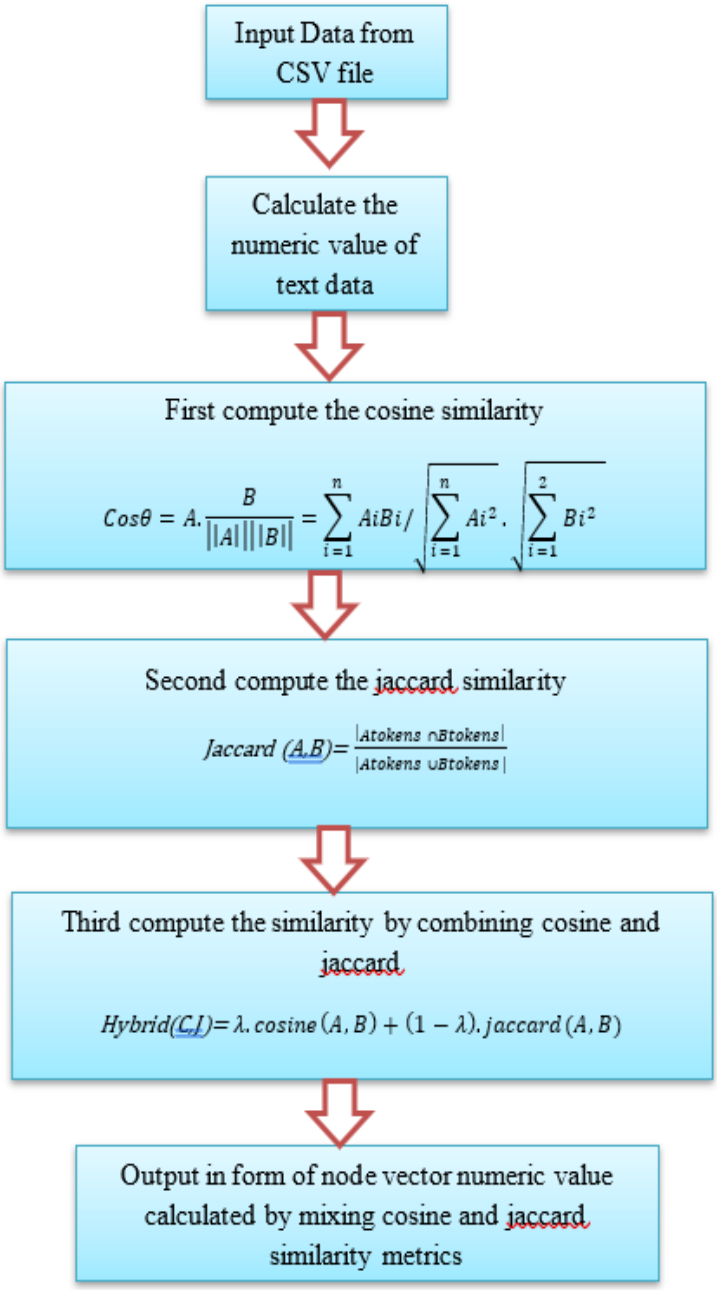


Figure 3. Workflow breakdown of Knowledge Graph Layer

3.3 Third Layer (Transformer Embedding layer)

The transformer has most important component is attention- states that 'attention is all you need' which comprises of self-attention, multihead attention, positional encoding. Taking the original transformer architecture as base architecture we also aim to design the architecture for our work point of view. Here in figure 2 the third layer of model is taking similarity matrix and visual semantic features from first layer and second layer. The positional embedding vector is generated from this matrix which gives position of these words respect to these integer numbers. Basically this layer is act as encoder as similarity matrix is encoded in to a embedding vector. The formula used by positional encoder is

$$PE(pos, 2i) = \sin(pos/1000^{2i/dmodel}) \quad (5)$$

$$PE(pos, 2i+1 = \cos(pos/1000^{2i/dmodel}) \quad (6)$$

The PE formula is generating the position of each ith value in form of sine waves and each ith +1 in form of cosine to generate sequence of alternate sine and cosine. The masked multi head attention is comprising of linear layer and SoftMax layer which generates text relate to these positional embeddings which also act as a decoder phase and at last output is generated in form of long text of the inputted image. The MLM uses the hidden state tensors of typical size 768/1024 and SoftMax layer generates the vocab size as per the requirement of output. The detailed architecture how transformer decoder will decode each node vector and image embeddings for generating multiline captions of an image is shown in figure 4.

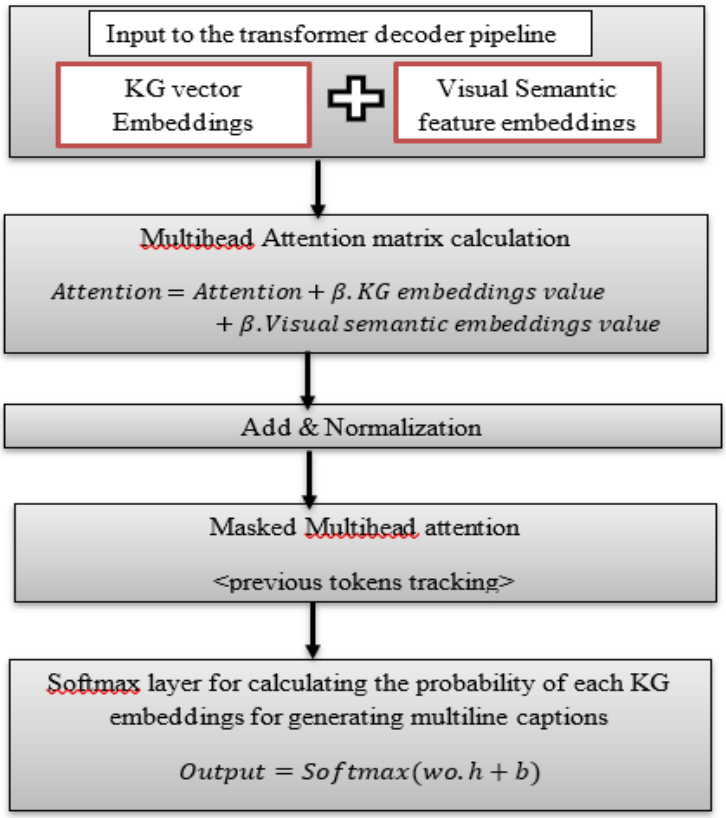


Figure 4. Working methodology of transformer decoder pipeline

The transformer decoder pipeline takes the KG vector and visual semantic feature embeddings for calculating the position vector of each embedding. The mutlihead attention matrix will generate the attention vector by using formula 7 which is the dot product of KG vector embeddings and visual feature embeddings. The attention matrix is normalized using Add & Normalization layer which is used to reduce the effect of vanishing gradient problem in decoding pipeline and stabilizes the training process by scaling the output across the features. The output in multiline summarization of image is generated by using SoftMax layer.

Attention=Attntion+β.KGembeddingvalue+β.visualsemanticembeddingsvalue (7)

Where

β=A weighting hyperparameter for controlling the contribution of embedding values to achieve the final attention value.

KGembeddingvalue=this component adds the domain understanding of textual features with matching semantics to image features.

Visualsemanticembeddingsvalue = this component helps in integrating the image features in textual attention.

Table 2. Algorithmic steps for Image caption Generation

Function feature extraction()
For each image in image data #calculate the feature extraction
Image feature=Resnet50(imagearray)
Imagearray=alexnetlayers(96,256,384,384,256)
returns image feature matrix
Function knowledgegraph(actualcaption)
For each token in image feature array # calculate vector
Tokens=text.split()
Vector=len(vocabulary)
For each vector in vector array # calculate the cosine similarity between vectors
Similarity score=cosine_similarity[vector1, vector2]
Similarity score=jaccard_similarity[vector1,vector2]
returns similarity score matrix
Function generatecaption(image features score, maxlength)
for each image in doc #Calculate captions as textual descriptions
captions=transformermodel.generatecaption= $PE(pos, 2i+1 = \cos(pos/1000^{2i/dmodel})$
return textual descriptions

In this study the dataset used is as Indiana university dataset which consist of 7420 images of chest xray. The dataset consist of two csv file as projections and radiology report in form of description of each image. The projection csv file consist of type of chest xray image as frontal or either lateral. The radiology report file consist of attribute like uid, mesh which is having classified values as normal, cardiomegaly, osteophyte, pulmonary, problems, indication, comparison, findings, impression**.** To handle the similar distribution for accurate output we divide the dataset by 70% training, 15%validation and 15% testing. For eg the dataset consist of 7420 images and 3955 description set of images, the split ration as per 70% is 2770 images for training a model, 15% for validation which consist of 590 images and another 15% for testing which consist of 590 images and split from 3955 textual description, 2765 as per 70% ratio for training set, 593 for validation and test set as per 15% ratio. The reason to take this dataset as it is publicly available on Kaggle and the format for image is jpeg not dicom. Other datasets are also available for chest xray but they are not handy to use as require lots of resources for training. The link for dataset is https://www.kaggle.com/datasets/raddar/chest-xrays-indiana-university is provided in table 3. The other datasets is listed as follows. To handle the class imbalance and to ensure the similar distribution of classes for proper subset of labels for each type of condition, we have utilized the 'findings' feature to categorize each findings with appropriate conditions for proper subset distribution during training, testing and validation phase. For eg the training, validation and testing set for some of the labels is depicted in table 4 summarized below.

4.2 Dataset Analysis

The dataset comprises of mesh attribute which consist of labels of abnormalities for each chest xray image. The problem attribute consist of specific abnormality to each image. The other attributes like findings, impression and indication consist of textual description for particular chest xray image. The mesh and problem attribute are correlated to each other so can be analysed with Heatmaps which is used plot correlation between attributes when dataset is too large. The dataset's finding column consist of description of each image but column has missing entries and description consist of noise terms like 'XXX' or "XX' which is not producing the accurate summary of an image. Here we develop an algorithm to clean the dataset by measuring the placeholder of 'XXX' or 'XX' with respect to each sentence context in form of procedure or diagnosis.

Table 3. Available dataset for chest xray radiology

Sr.No	Data set Name	Dataset Description	Dataset Availability Link
1	NIH chest xray dataset	It consists of 10,000 images of chest xray in PNG format where annotation file only consist of singe label to each image not whole description of image generated by radiologist	https://nihcc.app.box.com/v/ChestXray-NIHCC
2	Chest xrays	It consist of 5000 images in two classification modes as pneumonia and normal without annotation csv file	https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia
3	VinDrCXR	This dataset is largest dataset comprises of 18000 images with radiologist generated annotations but not freely available. Only the spine dataset is freely avaialble	-
4	Chexpert-chest xrays	This dataset is developed by standford university with 65240 images from October 2002 to July 2017. Dataset is not freely available	-

Table 4. Dataset stratified for training, validation and testing by using 70 % rule, 30 %rule (Val (15%) and test (15%))

Sr.No	Condition	Total	Train	Validation	Test
1	Cardiomegagly	134	94	20	20
2	Pleural Effuion	1409	986	211	211
3	Pneumothorax	1401	981	210	210
4	Atelectasis	221	155	33	33
5	Edema	188	132	28	28
6	Consolidation	1033	723	155	155
7	Fracture	69	48	10	10
8	Nodule	73	51	11	11
9	Opacity	370	259	56	56
10	Infiltration	0	0	0	0
11	Emphysema	43	30	7	7
12	Hypo inflation	2	1	0	0
13	Airspace disease	416	291	62	62
14	Lung patchy	1	1	0	0
15	Diaphragms	0	0	0	0
16	Bilateral	157	110	24	24
17	Pulmonary	1064	745	160	160
18	Spondylosis	59	41	9	9
19	Abdomen	55	39	8	8
20	Thoracic Vertebrae	0	0	0	0
21	Degenerative	452	316	68	68

Table 5. Algorithm for cleaning the dataset

Input- Indiana University Dataset(Excel file with ‘findings’ column)

Output- cleaned Dataset(CSV with conditionally cleaned ‘findings’)

1. Read the dataset using pandas library by calling the function readexcel() and obtained a Dataframe DF.
2. Remove rows where findings column entry is null or containing whitespace.
3. For each row in DF:

a. Let [sentences]=”findings”

b. If sentences contains “procedures’’ or ‘’diagnosis” followed by “XXX” or “XX” the Leave text unchanged

c. Else: Remove all standalone occurrences of the token “XXX” or “XX”

4. Store the results in new column as cleaned and save the results in CSV file for further processing

For eg if the substring ‘XXX’ or ‘XX’ appears after procedure or diagnosis then it is assumed that it is used for protecting health information or pending diagnosis otherwise the algorithm removes the standalone occurrences of ‘XXX’ or ‘XX’ is indicated in table 5.

5. Results and Discussion

The model is created using python language in spyder notebook where created three different files for capturing visual semantic feature of image, calculating the cosine distance between sentences of textual data and transformer model for encoder and decoder operations for summarization. The hyper parameter tuning parameters used for proposed model is shown in table 6.

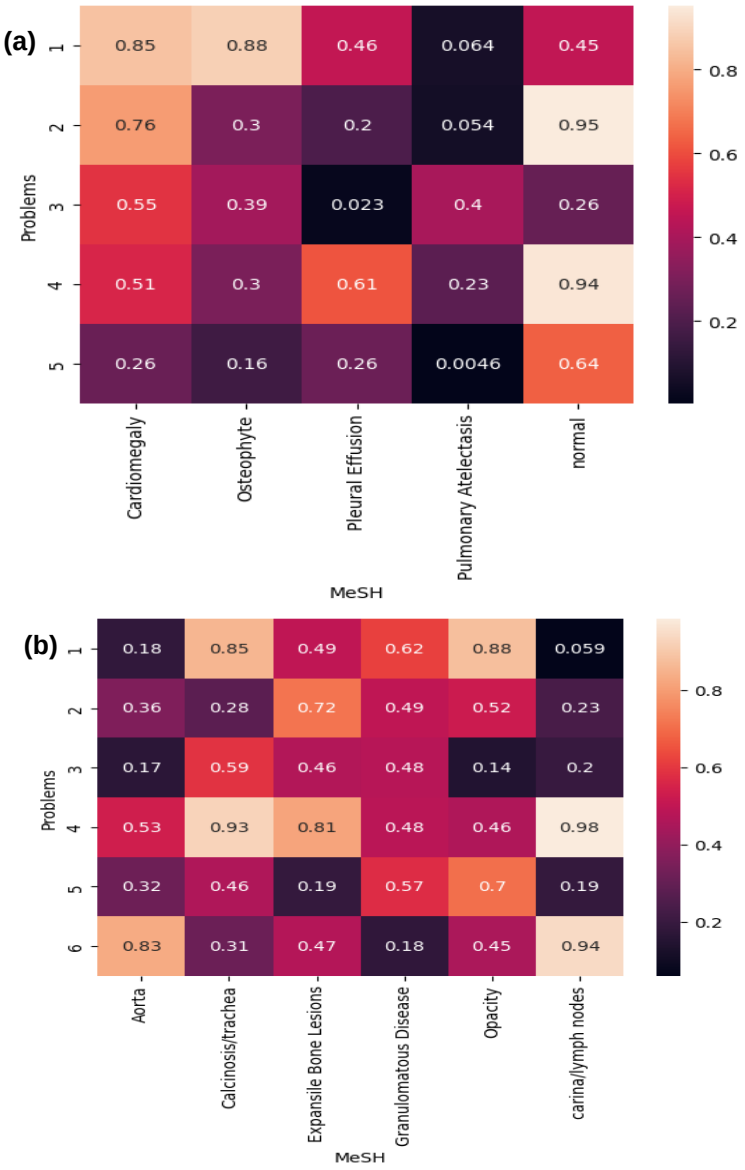


Figure 5. Correlated values of mesh conditions in dataset with respect to problems

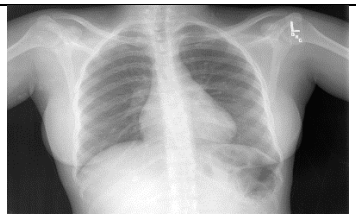
Table 6. Hyperparameter tuning

Parameter for Hybrid model(Alexnet+Resnet 50)	
Parameter	Values
Learning rate	[1e-4, 1e-3, 5e-3]
Optimizer	Adam
Batch size	64
Weight-decay	[1e-4, 1e-3]
Scheduler	['StepLR']
Epochs	500
Parameter for Knowledge Graph Layer	
Parameter	Values
λ	0.5 equal weight to cosine and jaccard
Parameter for Transformer layer	
Hidden size	1024
Num_hidden_layers	12
Dropout_rate	0.1 to 0.3
Max_position_embeddings	2048
Attention_droput	0.1
Warmup_steps	3000

5.1 Model Evaluation with other SOTA algorithms

The model is evaluated using different evaluation metric which measure the performance of proposed work. The explanation for BLEU-1, BLEU-4, F1, Precision, RIBES, METEOR, Spice and Recall is as follows. The dataset described above is undergone with four models 1) LSTM 2) BILSTM 3) Basic Conv2D and 4) Proposed Model. The proposed model is givng results better in evaluation metrics such as BLEU-1-63%, BLEU-2-61%, RIBES-79%, F-measure-85%, precision-85%, recall-83%, Spice-82% and METEOR-79% score as shown in table VII below. The proposed model is outperforming than baseline models is due to the introduction of more concrete graph structures as knowledge graph layer which provides the context awareness and factual consistency sentence embeddings to the transformer decoder for precise multiline captions of a medical image. The properties like object hallucination, loss evaluation mismatch and Bias is also improved.

Table 7. Outcome of proposed model with respect to ground truth

Input Image	
Ground Truth	Normal chest XXXX the cardiac silhouette & mediastinum size are within normal limits there is no pulmonary edema there is no focal cosoliation there are XXXX of a pleural effusion there is no evidence of pneumothorax
LSTM	no size evidence normal a no size evidence normal a no size evidence
BILSTM	Limits effusion size evidence no limits effusion size evidence no limits effusion size
Conv2D-AE	pulmonary no no silhouette no pulmonary no no silhouette no pulmonary
Proposed Model	The cardiac silhouette & mediastinum size are within normal limits there is no pulmonary edema there is no focal consolidation there is no pleural effusion there is no evidence of pneumothorax

1 Object hallucination is reduced by fusing KG embeddings with visual semantics features in to the attention layer, so model gives higher weight to context aware verified entities thereby not producing any entity not related to context.

2 Loss evaluation mismatch- The integrating of KG layer with transformer decoder improves semantic and factual coherence during training and correlates with evaluation metrics such as SPICE, METEOR and BLEU thereby aligning the captions with stated metrics despite of cross-entropy loss during training phase.

3 Bias- The transformer decoder pipeline by implementing formula 7 ensures that both factual knowledge and real visual cues are used thereby reducing reliance on biased correlations. The table VII shows the output obtained with the proposed model. The baseline models such as LSTM, BILSTM and Conv2d-AE suffer from redundancy and object hallucination which is solved with proposed model as the model obtained the multiline caption close to ground truth by removing the noisy substrings from the description.

Table 8. Evaluation Metrics Results between proposed model and other models

Evaluation Metrics	LSTM	BI LSTM	Conv2D-AE	Proposed Method
BLEU-1	0.54	0.47	0.36	**0.63**
BLEU-4	0.34	0.37	0.26	**0.61**
RIBES	0.36	0.41	0	**0.79**
F1	0.64	0.67	0.56	**0.85**
Precision	0.6	0.62	0.59	**0.85**
Recall	0.67	0.65	0.69	**0.83**
Spice	0.34	0.37	0.26	**0.82**
METEOR	0.34	0.21	0.16	**0.79**

In the below graph BLEU-1 stands out to be beter from other existing models. BLEU-1 is a specific variant of the BLEU(Bilingual Evaluation Understudy) metric that measure the precision of unigrams (individual words) in the generated text.This metric is usually used in natural processing & machine translation task to measure the quality of machine-generated text. The BLEU-1 score is calculated using the following components a)Precision-measures the percentage of unigrams available in the text. B)Bravity penality- is applied to avoid favoring overly short translations. This penalty adjusts the precision score based on the length of the genrated text compared to the reference text. The formula of BLEU-1 is stated as

$$BLEU - 1 = BP * \exp(1/N \sum_{n=1}^{n} logpn) \quad (8)$$

Bp-bravity penalty and N-is the maximum order of n-gram considered. Pn is the precision for n-gram range from 0 to 1. Higher BLEU-1 score generally

indicate goodness of an algorithm. The above graph gives BLEU-1 score for proposed method is 63% better than existing methods.

The bleu-4 metric specifically focuses on four grams, so it calculates precision based on the presence of four-word sequences

$$BLEU - 4 = BP * exp(\frac{1}{4\sum_{n=1}^{4} logpn}) \quad (9)$$

Where pn is the precision of four grams. The below graph in fig 7 explains about the method proposed method is analyzing well the four grams precision from the observed dataset than existing methods. The above graph is showing 61% better score than existing methods.

The F1score also know as the F1 measure or F1 score is a metric commonly used in the field of ML and statistics particularly in binary classification problem. The formula consist of harmonic mean between precision and recall and is often used when both false positives and false negatives are important. The formula as stated as

$$F1 = 2 * precision * recall/precision + recall \quad (10)$$

Here, precision is the count of true positive predictions divided by the total number of positive predictions(true positive + false positives) & recall or sensitivity is the count of true positives predictions divided by the total number of actual positives (true positives + false negatives). The F1 score range from 0 to 1 where 1 determines perfect precision & recall with 0 indicates the worst possible. The below graph in fig 7 is giving 85% percent better than existing methods.

The recall is the senstivity or true positive rate metric used in the evaluation of calssification models. It masures the ability of a model to capture all the relevant instances of a positive class. The formula is stated as

$$Recall = \frac{True\ positive}{True\ positive + False\ negatives} \quad (11)$$

A perfect value of recall value indicates that model is effective at capturing the positive instances effectively. Recall is not using alone, used alongside with precision. The below graph in fig 6 showing recall value of 83% better than existing methods.

The precision is one of the evaluation metric used for evaluation of classification models. It represents the ratio of true positive predictions to the total number of positive predictions made by the model. It focuses on the accuracy of positive predictions is defined by the following formula

$$Precision = \frac{True\ positive}{True\ positive + False\ negatives} \quad (12)$$

A high precision value means that the model has a lower rate of false positive result, it is likely to be correct. The below graph in fig 7 is showing the precision measure of 85 % better than existing methods.

Semantically propositional image caption evaluation, or SPICE for short, is a metric commonly used to evaluate the quality of generated captions based on semantic propositions in the context of picture captioning. The spice value in the graph above is 82%, but the current methods only draw up to 20%, 30%, and 35%, respectively. The suggested strategy accurately predicts the quality of the image captions.

The RIBES stands for rank-based intuitive bilingual evaluation score is a metric used for the evaluation of machine translation output. It's designed to address some of the limitations of other metrics like BLEU by considering the relative ranks of words instead of exact matches. It calculates the score with three formula a)precision b) geometric mean and c) normalization. Precision calculation formula is stated

$$pi = \frac{Number\ of\ words\ at\ position\ i\ in\ both\ candidate\ and\ reference}{Number\ of\ words\ at\ postion\ i\ in\ the\ candidate} \quad (13)$$

Here candidate and reference is the total number of different words in the dataset with respect to sentences. Geometric mean is precision score at different words. formula is stated as

$$G = (\prod \sum_{i=1}^{N} pi)^{1/N} \quad (14)$$

Where N is the total number of word positions in the translations. Normalization mean by a factor to account for variations in translation length. The formula is stated as

$$RIBES = \frac{G}{Normalization\ Facor} \quad (15)$$

It often based on the length of the reference translation. The below graph in figure 6 is showing 79% better result compare to existing methods also shown in figure 7.

The METEOR is a metric for evaluation of translation with explicit ordering is an automatic evaluation metric used to assess the quality of machine-generated translations. It was designed as an alternative or complement to other metrics like BLEU and TER (Translation Edit Rate). It considers various factors such as precision, recall, stemming, synonymy and word order. The formula is stated as

$$METEOR = (1 - \mathcal{B}) * precision + \mathcal{B} * recall \quad (16)$$

Where β is a tunable parameter that determines the balance between precision & recall. The actual precision & recall value involves contribution from unigrams, synonymy, and stemming and word order components.

5.2 Quantitative Analysis of proposed model

Here we performed the analysis of variance test to check the hypothesis between the models. To check the mean differences between the proposed model and other models we utilized the one form of ANOVA i.e single factor ANOVA.

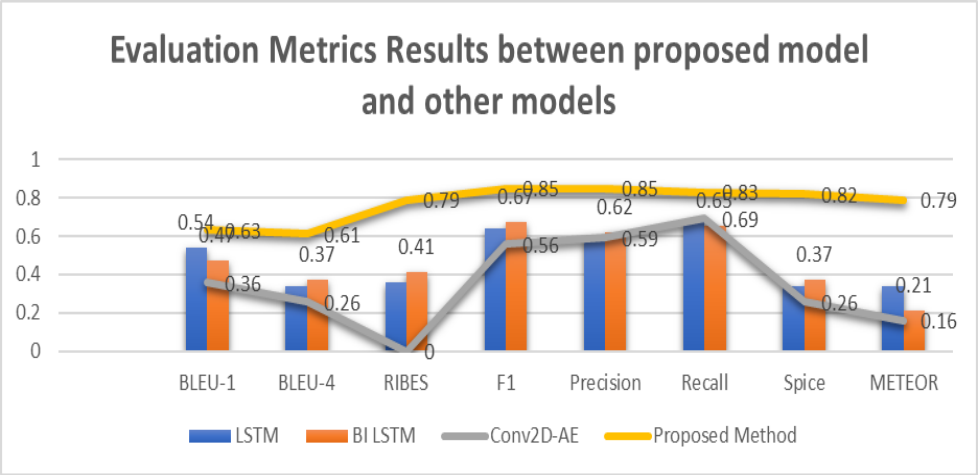


Figure 6. Comparison between proposed model and other state of art algorithm

Table 9. Annova Test on models

Source of Variation	SS	df	Ms	F	P-value	F critical
Between Groups	0.136533	1	0.136533	4.109562	0.070131	4.964603
Within Groups	0.332233	10	0.033223			
Total	0.468767	11				

Table 10. Comparison between Proposed model and other previously implemented models

Sr. No	Model Name	BLEU-1	BLEU-4
1	Densenet+KG	0.44 [4]	0.14 [4]
2	EDCnet	0.51 [7]	0.17 [7]
3	Full ARL	0.12 [8]	0.17 [8]
4	Relation Paranet	0.50 [9]	0.17 [9]
5	TrMRG	0.53 [10]	0.15 [10]
6	MDINAP	0.38 [11]	0.12 [11]
7	Proposed Model (Hybrid resnet+Alexnet,KG, transforomer)	**0.63**	**0.61**

To check the variability two hypothesis is carried out a) Null hypothesis- when there is no difference between mean values of observed groups. b) Alternative hypothesis- when there is difference between mean values of observed groups. Here in the given figure the p value is 0.07 which is compared with significance value is 0.5. The significance of hypothesis is measured with the help of p-value which is the probability of observing the obtained data assuming the null hypothesis is true. A small p value suggests that the observed data is unlikely under H0, indicating stronger evidence against it. If p value is less than 0.5 then reject the null hypothesis and accept the alternate hypothesis. If p value is greater than 0.5 then accept the null hypothesis. Here we reject the null hypothesis as there is difference in mean between existing model and our proposed model as shown in below table IX.

5.3 Comparison between Proposed model and other models cited in related work section

The below graph and above the table are showing the increment in result in our proposed model and other model which implemented early in shown in table X. Our model is giving better results in terms of BLEU-1 and BLEU-4 score carried out on same dataset that is Indiana university dataset. The results turnout to be better as we introduced knowledge graph layer between hybrid resnet50 & alexnet model and transformer.

6. Ablation Studies

The proposed is integrating with hybrid vision algorthims as alexnet and Resnet 50, knowledge graph layer and transfromer layer which improves the multiline caption of an understudy medical image.

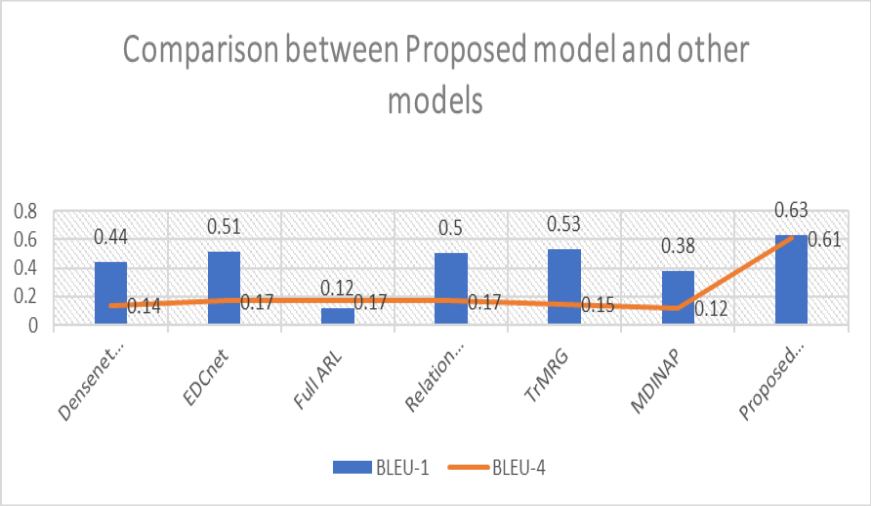


Figure 7. Comparison graph between models

The drop of first layer reduces the performance of model by 5 % in terms of evaluation metrics. The drop of second layer impacts the model perfomance by 25 % in terms of each evaluation metrics. The drop of transfromer layer reduces the model performance by 5%. After removing the each component from full model the value of evaluation of evaluation metrics is reduced to BLEU-1- 0.28, BLEU-4- 0.26, RIBES-0.44,F1- 0.5, precison-0.5, Recall- 0.43, SPICE- 0.47, METEOR-0.44.

7. Conclusion

In this study we have used Indiana university dataset of chest Xray to generate the automatic image captioning but in long text description. Existing model like BILSTM, LSTM and Conv2DAE is just generate only short text description of an inputted image. To solve this issue, we have proposed a novel framework based on knowledge graph layer between the first stage as Hybrid model of computer vision deep learning algorithms and third stage as transformer-based model. It is analyzed from results that proposed model is outperform than existing models in terms of BLEU-1-63%, BLEU-4-61,RIBES-79%,F-1-85%,Precision-85%,Recall-83%, Spice-82% and METEOR- 72%. so, in future work first task is to remove noise for smooth generation of summary without using the annotated file and generate an improvised dataset which still is a challenge for multi modal summarization task. There is broader challenges exist in multimodality problems as 1) The models trained on one medical domain often poorly performs on other medical datasets as pathology reports, dermatology reports etc. 2) Aligning visual and textual information is a difficult task as sometime text may not describe all parts of an image due to the limitations in the annotated file so open research work is going on obtaining the dynamic knowledge grounding for automatic annotation capturing. These challenges help the future researcher to work in the medical domain for improving the healthcare data summarization domain.

References

[1] T. Ghandi, H. Pourreza, H. Mahyar, Deep learning approaches on image captioning: A review. ACM Computing Surveys, 56(3), (2023) 1-39. https://doi.org/10.1145/3617592

[2] Y. Lin, K. Lai, W. Chang, Skin medical image captioning using multi-label classification and Siamese network. IEEE Access, 11, (2023) 23447-54. https://doi.org/10.1109/ACCESS.2023.3249462

[3] J.H. Moon, H. Lee, W. Shin, Y.H. Kim, E. Choi, Multi-modal understanding and generation for medical images and text via vision-language pre-training. IEEE Journal of Biomedical and Health Informatics, 26(12), (2022) 6070-6080. https://doi.org/10.1109/JBHI.2022.3207502

[4] Z. Wang, H. Han, L. Wang, X. Li, L. Zhou, Automated radiographic report generation purely on transformer: A multicriteria supervised approach. IEEE Transactions on Medical Imaging, 41(10), (2022) 2803-13. https://doi.org/10.1109/TMI.2022.3171661

[5] Y. Zhang, X. Wang, Z. Xu, Q. Yu, A. Yuille, D. Xu, When radiology report generation meets knowledge graph. InProceedings of the AAAI conference on artificial intelligence, 34(7), (2020) 12910-12917. https://doi.org/10.1609/aaai.v34i07.6989

[6] Y. Peng, Y. Tang, S. Lee, Y. Zhu, R.M. Summers, Z. Lu, COVID 19 CT CXR: a freely accessible and weakly labelled chest X ray and CT image collection on COVID 19 from biomedical literature. IEEE Transactions on Big Data, 7(1), (2020) 3 12. https://doi.org/10.1109/TBDATA.2020.3035935

[7] D. Singh, M. Kaur, J.M. Alanazi, A.A. AlZubi, H.N. Lee, Efficient Evolving Deep Ensemble Medical Image Captioning Network. IEEE Journal of Biomedical and Health Informatics, 27(2), (2023)

1016–25. https://doi.org/10.1109/JBHI.2022.3223181

[8] D. Hou, Z. Zhao, Y. Liu, F. Chang, S. Hu, Automatic report generation for chest X ray images via adversarial reinforcement learning. IEEE Access, 9, (2021) 21236–21250. https://doi.org/10.1109/ACCESS.2021.3056175

[9] F. Wang, X. Liang, L. Xu, L. Lin, Unifying relational sentence generation and retrieval for medical image report composition. IEEE Transactions on Cybermetrics, 52(6), (2020) 5015–5025. https://doi.org/10.1109/TCYB.2020.3026098

[10] M.M. Mohsan, M.U. Akram, G. Rasool, N.S. Alghamdi, M.A.A. Baqai, M. Abbas, Vision Transformer and Language Model Based Radiology Report Generation. IEEE Access, 11, (2023) 1814–1824. https://doi.org/10.1109/ACCESS.2022.32327

[11] H. Park, K. Kim, S. Park, J. Choi, Medical image captioning model to convey more details: Methodological comparison of feature difference generation. IEEE Access, 9, (2021) 150560–150568.

[12] W. Wang, R. Wang, X. Chen, (2021) Topic scene graph generation by attention distillation from caption. In Proceedings of the IEEE/CVF international conference on computer vision, IEEE, Montreal, QC, Canada, 15900–15910. https://doi.org/10.1109/ICCV48922.2021.01560

[13] P. Qi, Z. Huang, Y. Sun, H. Luo, (2022) A Knowledge Graph Based Abstractive Model Integrating Semantic and Structural Information for Summarizing Chinese Meetings. In Proceedings IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD), IEEE, Hangzhou, China, 746–751. https://doi.org/10.1109/CSCWD54268.2022.9776298

[14] J. Guo, Y. Wang, (2021) Summarizing RDF graphs using Node Importance and Query History. In Proceedings IEEE 2021 International Conference on Service Science (ICSS),IEEE, Xi'an, China, https://doi.org/10.1109/ICSS53362.2021.0001

[15] M. Aamir, A.U. Jan, N. Mukhtar, M.A. Khan, Z. Ali, W.A. Abro, Y. Guan, An unsupervised graph-based hybrid approach for opinion summarization. In Proceedings 18th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), IEEE, Chengdu, China, 83–88. https://doi.org/10.1109/ICCWAMTIP53232.2021.9674086

[16] H. Zhan, K. Zhang, C. Hu, V.S. Sheng, (2021) Gated Graph Neural Networks (GG NNs) for Abstractive Multi Comment Summarization. In Proceeding IEEE Int Conf Big Knowledge (ICBK), IEEE, Auckland, New Zealand, 323–330. https://doi.org/10.1109/ICKG52313.2021.00050

[17] U. Barman, V. Barman, M. Rahman, N.K. Choudhury, Graph based extractive news articles summarization approach leveraging static word embeddings. In Proceedings 2021 International Conference on Computational Performance Evaluation (ComPE), IEEE, Shillong, India, 8–11. https://doi.org/10.1109/ComPE53109.2021.9752056

[18] R. Jalota, D. Vollmers, D. Moussallem, A.C.N. Ngomo. (2021) LAUREN – Knowledge Graph Summarization for Question Answering. In Proceeding IEEE 15th International Conference on Semantic Computing (ICSC),IEEE, Laguna Hills, CA, USA, 221–226. https://doi.org/10.1109/ICSC50631.2021.00047

[19] E. Yang, F. Hao, J. Gao, Y. Wu, G. Min, (2020) Entity spatio temporal evolution summarization in knowledge graphs. In 2020 IEEE International Conference on Knowledge Graph (ICKG), IEEE, Nanjing, China, 181–187. https://doi.org/10.1109/ICBK50248.2020.00035

[20] T. Yao, Y. Pan, Y. Li, T. Mei, (2019) Hierarchy Parsing for Image Captioning. In Proceeding IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Seoul, Korea (South), 2621–2629. https://doi.org/10.1109/ICCV.2019.00271

[21] A. Jangra, S. Mukherjee, A. Jatowt, S. Saha, M. Hasanuzzaman, A survey on multi modal summarization. ACM Computing Surveys, 55(13s), (2023) 1-36. https://doi.org/10.1145/3584700

[22] S.K. Uppada, P. Patel, B. Sivaselvan, An image and text based multimodal model for detecting fake news in OSN's. Journal of Intelligent Information Systems, 61(2), (2023) 367–393. https://doi.org/10.1007/s10844-022-00764-y

[23] B. He, J. Wang, J. Qiu, T. Bui, A. Shrivastava, Z. Wang, (2023) Align and Attend: Multimodal Summarization with Dual Contrastive Losses. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, IEEE, Vancouver, BC, Canada. https://doi.org/10.1109/CVPR52729.2023.01428

[24] M. Xiao, J. Zhu, H. Lin, Y. Zhou, C. Zong, (2023) CFSum: A Coarse-to-Fine Contribution Network for Multimodal Summarization. arXiv preprint arXiv:2307.02716. https://doi.org/10.48550/arXiv.2307.02716

[25] T. Gigant, F. Dufaux, C. Guinaudeau, M. Décombas, (2023) TIB: A Dataset for Abstractive Summarization of Long Multimodal Videoconference Records. In Proceedings of the 20th International Conference on Content-based Multimedia Indexing, 61-70. https://doi.org/10.1145/3617233.3617238

[26] J. Li, X. Wang, Y. Zhu, Y. Zhang, J. Tang, Elastic deep multi-view autoencoder with .0 diversity embedding. Neurocomputing, 2022. 512 41. https://doi.org/10.1016/j.neucom.2022.09.001

[27] D. Jha, S. Saha, N. Dey, Automatic colorectal cancer detection using machine learning and deep learning based on feature selection in histopathological images. Applied Soft Computing, 112, (2021) 107813. https://doi.org/10.1016/j.asoc.2021.107813

[28] Z. Wang, Y. Liu, X. Hu, Image captioning by diffusion models: a survey. Information Fusion, 93, (2023) 130–145. https://doi.org/10.1016/j.inffus.2023.04.002

Authors Contribution Statement

Both the authors equally contributed and approved the final version of this work.

Funding

The authors declare that no funds, grants or any other support were received during the preparation of this manuscript.

Competing Interests

The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

Data Availability

The data supporting the findings of this study can be obtained from the corresponding author upon reasonable request.

Has this article screened for similarity?

Yes

About the License