IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 6 December 2022, accepted 14 December 2022, date of publication 27 December 2022, date of current version 6 January 2023.

Digital Object Identifier 10.1109/ACCESS.2022.3232719

RESEARCH ARTICLE

Vision Transformer and Language Model Based Radiology Report Generation

MASHOOD MOHAMMAD MOHSAN[1], MUHAMMAD USMAN AKRAM[1], (Senior Member, IEEE), GHULAM RASOOL[2], (Senior Member, IEEE), NORAH SALEH ALGHAMDI[3], MUHAMMAD ABDULLAH AAMER BAQAI[4], AND MUHAMMAD ABBAS[1]

[1]Department of Computer and Software Engineering, National University of Sciences and Technology, Islamabad 44000, Pakistan
[2]Machine Learning Department, Moffitt Cancer Center, Tampa, FL 33612, USA
[3]Department of Computer Sciences, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, Riyadh 11671, Saudi Arabia
[4]College of Engineering, Michigan State University, East Lansing, MI 48824, USA

Corresponding authors: Mashood M. Mohsan (mashood3624@gmail.com) and Norah Saleh Alghamdi (nosalghamdi@pnu.edu.sa)

This work was supported by the Princess Nourah bint Abdulrahman University Researchers Supporting under Project PNURSP2023R04, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

ABSTRACT Recent advancements in transformers exploited computer vision problems which results in state-of-the-art models. Transformer-based models in various sequence prediction tasks such as language translation, sentiment classification, and caption generation have shown remarkable performance. Auto report generation scenarios in medical imaging through caption generation models is one of the applied scenarios for language models and have strong social impact. In these models, convolution neural networks have been used as encoder to gain spatial information and recurrent neural networks are used as decoder to generate caption or medical report. However, using transformer architecture as encoder and decoder in caption or report writing task is still unexplored. In this research, we explored the effect of losing spatial biasness information in encoder by using pre-trained vanilla image transformer architecture and combine it with different pre-trained language transformers as decoder. In order to evaluate the proposed methodology, the Indiana University Chest X-Rays dataset is used where ablation study is also conducted with respect to different evaluations. The comparative analysis shows that the proposed methodology has represented remarkable performance when compared with existing techniques in terms of different performance parameters.

INDEX TERMS Vision transformers, language models, radiology report, decoder.

I. INTRODUCTION

Diseases that target the chest are particularly dangerous because of their impact on the lungs. Every year, millions of people face being diagnosed with a chest disease [1]. As the lungs are important organs in the human body, any damage caused to them could have life threatening implications. On average, 58000 deaths occur only due to pneumonia [2]. Chest x-rays are the primary practice to diagnose these diseases. The radiologist conducts a thorough visual of the x-ray image and writes a report of the patient's condition. This implies the same definition of image based report generation which is the task of describing the visual content of image in natural language by understanding the visual semantics [3].

The associate editor coordinating the review of this manuscript and approving it for publication was Carmelo Militello.

A manually created report describes the general chest condition, findings, and diseases if found. The radiologist must possess the following skills to correctly read a chest x-ray [3]: thorough knowledge of the basic anatomy of thorax as well as physiology various chest diseases, ability to analyze the radiograph through identifying different pattern, ability to analyze and evaluate the evolution over time of chest x-rays and recognize any changes that might occur, knowledge of clinical presentation and history, Knowledge of correlation to diagnostic results. This laborious task can result in being error prone if written by an inexperienced physician while simultaneously being tedious and time consuming for an experienced physician. The problem of report generation being a lengthy process is highlighted when large amounts of chest x-rays need to be analyzed. In highly populated areas, there could be hundreds of chest x-rays to analyze daily. Even after gaining

extensive experience, it takes a radiologist, on average, 5 to 10 minutes to correctly read a chest x-ray [4].

Generating a diagnostic report is actually a image-to-sequence problem whose inputs are pixels. A complete diagnostic report consists of findings, impressions, and tags. Previous solutions make use of a multi- tier system. The tags are considered as labels and a multi class classifications produces predicted labels for each chest x-ray image [4]. After performing a semantic analysis on the image, the correct description is attached. Descriptions in the reports are multi sentence long and their generations is crucial to the accuracy and quality of the report.Many solutions employing LSTM network has been proposed to solve this problem such as [4]. However, report descriptions consist of long sentences and CNN fails to encode complete features in latent space therefore effecting the accuracy's of report generated by LSTM.

The literature reports earlier attempts to create a radiologist report for a chest X-ray image by incorporating multiple CNN-RNN frameworks. In 2017, researchers proposed Transformer, a new simple network architecture that completely eschews recurrence and convolutions in favour of attention mechanisms. Convolutional or recurrent neural networks in an encoder-decoder arrangement constitute the foundation of the earlier dominating sequence transduction models. The Transformer model also uses the same configuration but relies only on attention mechanism. In order to increase overall efficiency, this article provides a novel Transformer Medical Report Generator (TrMRG) method for producing a chest X-ray report.

The contributions of this paper are summarized as follows:

1) We propose TrMRG (Transformer Medical report generator), an end-to-end Transformer-based model for report generation with pre-trained computer vision (CV) and language models. Although TrOCR was the first to adopt this architecture but it was used for only classification purposes. To the best of our knowledge, this is the first effort to use pre-trained image and text Transformers in tandem to generate medical reports.
2) A detailed ablation study has been conducted to identify best pre-trained models for encoder and decoder to generate more reliable reports
3) TrMRG achieves remarkable score with a standard Transformer-based encoder-decoder model, which is convolution and recurrence free and does not rely on training from scratch. It can be easily fine-tuned to predict accurate reports. The model will be made publicly available.

The remaining of the paper is structured as follows. Section 2 examines related literature. The methodology is explained in Section 3. The ablation studies and experimental results are presented in Section 4, and Section 5 contains the conclusion.

II. LITERATURE REVIEW

Automated medical report generation is the application of computer vision and language models which has strong societal impact. Medical report generation process started from [5] who proposed a CNN-RNN architecture to generate captions for images. These results were however too simple and lacked details. As more work was done in the field, attention was introduced and models like [6] used attention with RNN and CNN. This model produced significantly better results. Transformers were introduced in 2019 [7]. It is free from convolution and recurrence and solely focuses on attention. It used Multi attention self-attention (MSA) with multiple encoder and decoder layers which made it outperform previous language models. Since 2019, transformers are not only used with text but also with images where they have outperformed many existing techniques at different tasks. Here, we have divided the literature review into different subsection for better understanding.

A. TRANSFORMER IN LANGUAGE

Transformer models have performed admirably on a variety of linguistic tasks. BERT (Bidirectional Encoder Representations from Transformers) [8], GPT2 (Generative Pre-trained Transformer) [9], and T5 (Text-to-Text Transfer Transformer) [10] are a few of the well-liked models. BERT pre-trained the Transformer in a self-supervised manner using the Masked Language Model (MLM) and Next Sentence Prediction (NSP). In the Masked Language Model, 15% of the words in a sentence are randomly masked, after which the model is trained to predict these words using cross-entropy loss. The model gains the ability to take into account the bidirectional context while making predictions. The model predicts a binary label for a pair of sentences in NSP. The model may construct associations between two sentences as a result. They are essential in issues involving natural language, like question-answering and natural language inference. A different training approach was used in GPT-2, known as Causal Language Modelling, in which the model is trained to predict the next word given all the previous words. GPT-2 was stacked with only decoder layers of Transformer and token embedding along with positional embedings were calculated and added to sequences of tokens as input. Multi-head self-attention, feedforward network, layer normalization, and residual connections are applied by each layer. DistilGPT-2 [11] is a popular compressed version of GPT-2. It has fewer parameters compared to GPT2. DistilGPT-2 is trained on OpenWebTextCorpus by using Knowledge distillation methods. MiniLM [12] uses the same tokenizer as XLM-R while having its architecture based on BERT. To make the teacher's self-attention module more moldable and effective for the learner, final layer of the transformer is distilled.

B. TRANSFORMER IN IMAGES

Transformers were first used in 2018 [24] as Image generative model. In 2020, [25] Vision Transformer (ViT), also

TABLE 1. Literature review summary table.

Sr#	Method	Year	Approach
1	RTMIC [13]	2019	DenseNet_121(ROI) and Transformer(3 layer)
2	KERP [14]	2019	CNN and Graph + Transformer
3	BASE+RM+MCLN [15]	2020	Memory driven Transformer
4	RGNet [16]	2020	Two CNN and Two Transformer
5	KGAE-Supervised [17]	2021	Knowledge driven Transformer
6	MV+T+I [18]	2021	Multiple transformers architecture
7	PPKED [19]	2021	Distillation of posterior and prior knowledge via transformer
8	M^2triprogressive [20]	2021	DenseNet(CNN) - ViLM (Meshed memory Transformer) and BART
9	CDGPT2 [21]	2021	Chexnet + GPT2
10	AlignTransformer [22]	2022	Align hierarchy attention and Multi-grained Transformer used
11	Trust it or Not [23]	2022	Auto-encoder + transformer

known as vanilla image transformer, was proposed to demonstrate Transformer in image classification which outperformed existing image recognition benchmarks (ImageNet, CIFAR-100, VTAB, etc.). Vision Transformer (ViT) is the first implementation of a transformer in a deep neural network on large-scale image datasets. They took sequences of image patches which were flattened as vectors and then applied the original transformer model. The model was trained on a large dataset and then refined to downstream recognition benchmarks such as ImageNet classification. Spatial biasness is one of the inductive biasness present in CNN hence a CNN assumes a certain structure is present in images so it updates it filter parameters accordingly in order to classify images. Transformers does not inherit spatial biasness property opposed to CNN as they are solely based on attentions. DeiT [26] showed that it was possible to learn Transformers on mid-sized datasets in relatively shorter training episodes. It used procedures found in CNNs such as augmentation and regularization and adopted a unique knowledge distillation approach to train Transformers. A CNN model was employed as teacher model to distill a student Transformer. New properties to ViT that are different compared to convolutional networks were added by the DINO work [27]. This used self-supervised learning techniques. The term DINO is a method interpreted as a form of self-distillation with no labels. It provided solid foundation for the idea that self-supervised learning could be key to developing a ViT based BERT like model. The next model is ViT-MAE [28] which uses a simple MLM-like architecture. Patches of input images forming a series are given as input and majority are masked out. The remaining visible patches are then given the encoder. The encoded images along with MASK tokens added to them are given input to a decoder hence to reconstruct the original image. After pre-training is complete, the encoder is utilised for recognition tasks on un-corrupted images instead of the decoder, which is then destroyed. Table 2 shows are summary of different transformers.

TABLE 2. Comparison of different Image Transformers.

Sr#	Model	Dataset	Base	Large
1	Vit	IN-1K	77.9	76.53
2	DEIT	IN-1K	**85.2**	-
3	Dino	IN-1K	82.8	-
4	MAE	IN-1K	83.6	**85.9**

C. MEDICAL REPORT GENERATION

Automated medical report generation has a journey where different techniques have been utilized before transformed for this purpose. Early methods for writing reports included template filling, description retrieval, and hand crafted natural language generating techniques. In its basic arrangement, the task can be named as an image-to-sequence problem with pixel-values in form of series as inputs. These input patches of input are represented as feature vectors in the visual encoding stage, which encodes the input, also known as latent space vector, for the next generative step known as the language generation. A string of words or subwords that have been decoded using a particular vocabulary are the result of this. Xiong et al. [13] proposed a 2 part model. The first part was an Image Encoder and the second part was a non-recurrent Captioning Decoder. Li et al. [14] proposed the KERP (Knowledge-driven Encode, Retrieve, Paraphrase) approach that integrated contemporary learning-based methodologies for report generation with knowledge and retrieval-based methods. The work of producing reports was divided by KERP into learning medical abnormality graphs and subsequent natural language modelling. First, visual features were converted into a structured abnormality graph using an encoder. The templates were then obtained by a retrieve module based on the abnormalities found. A paraphrase unit then revised the templates to fit the given situation. A memory-driven transformer to produce a report was suggested by [15]. Their strategy involved equipping the transformer with a relational memory to store important data. Additionally, a memory-driven conditional layer normalisation was used to include memory into the transformer's decoder. Srinivasan et al. [16] proposed a deep neural network which predicted tags and created a report for a given chest x-ray. To get the tag embeddings, they used a convolutional neural network followed by transformers for learning self and cross attention. Image tags and features are encoded with self-attention to get a more detailed representation. Both of the above features were made use of in cross-attention, also known as Encoder-Decoder attention, along the sequence of input in order to generate the report

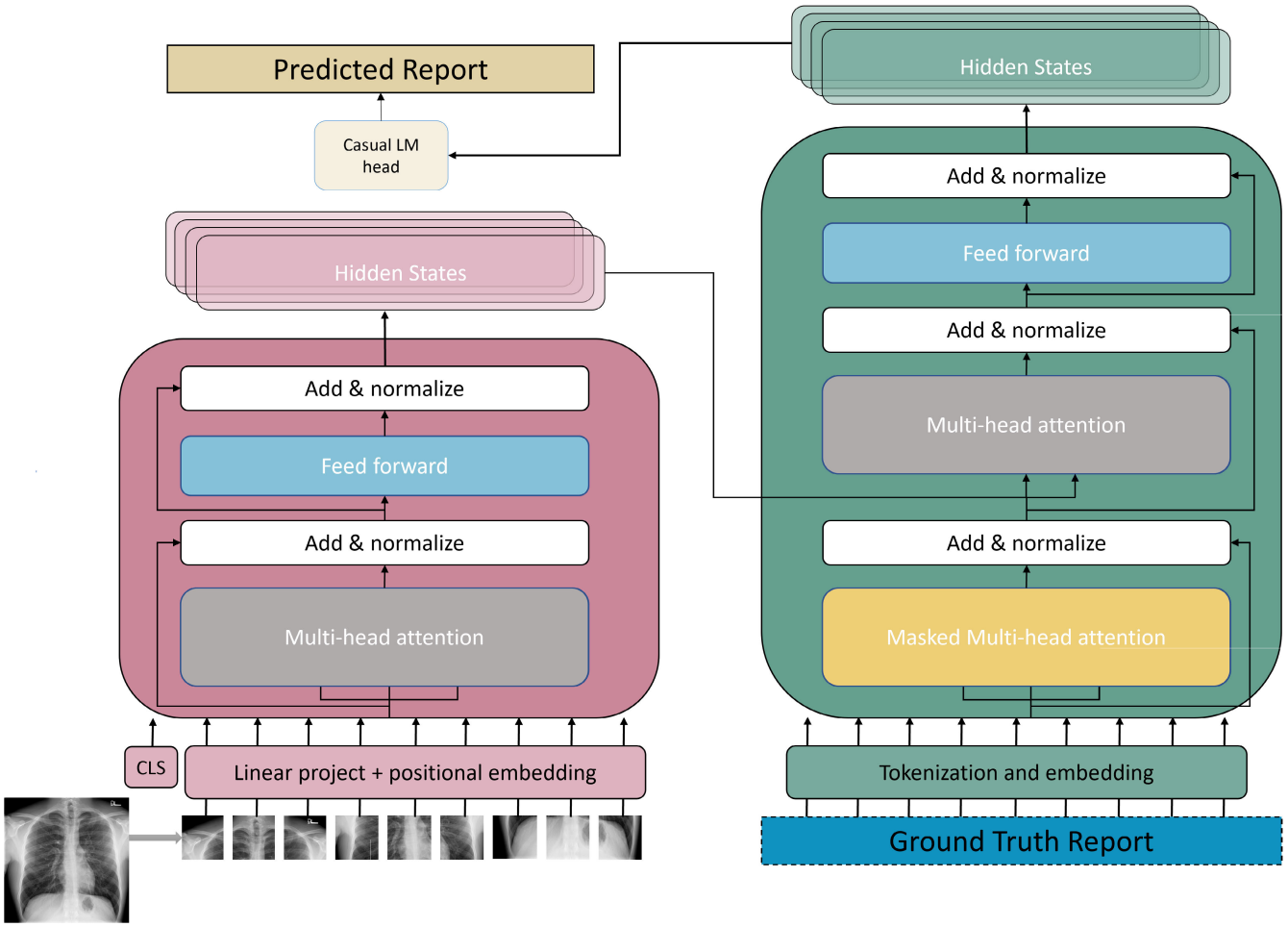


FIGURE 1. Illustration of proposed method. The left part shows the image based encoder which consists of N layers each containing attention heads. The right part shows the language based decoder consisting of N layers.

findings section. Impressions were generated by applying cross-attention on findings and input sequence. Reference [29] presented a method where they combined CNN based features with attention layer and LSTM to generate more reliable reports. They utilized IU and MIMIC datasets for evaluation purposes and outperformed simple attention based methods. A Knowledge Graph Auto-Encoder (KGAE) model that allowed unrelated sets of images and reports for training was proposed by [17]. An encoder and decoder driven by knowledge along with knowledge graph was also included in it. The visual and written realms were linked by the knowledge graph (like a latent space). The images were projected at the proper coordinates in the latent space where the encoder reported. The decoder used the provided coordinates to generate the report. This model was unsupervised because it may be trained using diverse sets of photos and reports. Nguyen et al. [18] proposed Classification of Clinical history and Chest X-ray to generate embedding of diseases along with a Transformer decoder sub-modules in an a fully differentiable paradigm to generate complete diagnostic reports. To ensure consistency with disease related topics, a weighted embedding representation was fed to the interpreter. Liu et al. [19] proposed a model which mimics the process of radiologists. First they identified the disease from the image using a CNN and then used prior knowledge for report generation. Nooralahzadeh et al. [20] proposed a two-step model which derived global concepts from the image then reformed them into finer and coherent texts using a transformer architecture. You et al. [22] proposed a AlignTransformer framework, Align Hierarchical Attention (AHA) and Multi-Grained Transformer (MGT) were the components of the AlignTransformer framework. Wang et al. [23] proposed a method which explicitly quantified visual and textual uncertainties for radiology report generation. Alfarghaly et al. [21] proposed a deep learning model consisting of CNN model as encoder and a Transformer model as decoder. They used Chexnet, as encoder, to predict the tags for images and also to generate latent space vector. Chexnet is one of the largest models used in this area. They then calculated weighted semantic features from the predicted tags pretrained embeddings. Finally to generate a report, they used a GPT2 pre trained model on the latent space vetor and semantic features.

GPT2 was one of the largest language generation model and despite this, their unique deep learning farmework underperformed when compared to others as their BLEU-1 score was only 0.387 and BLEU-2 score was 0.245. An end-to-end Transformer model for Optical Character Recognition was presented in [30]. It was the first research that utilized pre-trained weights for both image and language models. TrOCR was limited for being a classification model only. Table 1 shows an overview of literature review with respect to different methodologies their respective approaches.

III. METHODOLOGY

Fig 1 shows an overview of our proposed methodology. An encoder and a decoder are both parts of the TrMRG model. The encoder receives an image as input, breaks it up into patches, and then adds positional encoding to it. Each layer of encoder contains multiple self-attention heads which encode its feature representation. The encoded features are passed to multiple self attention heads in decoder layers in form of Queries Q and Keys K to decode it. After passing the output of the decoder to the linear layer and softmax, words probabilities are predicted. The output of encoder and decoder is stated as Hidden states in Fig 1. In literature, the output of encoder is also named latent space.

A. POSITIONAL ENCODING

Transformers by nature does not contain any recurrence or any convolution so in order to maintain order of sequence, position information must be stored. Tokens converted into embedding vectors are summed up with Positional Encoding vectors. For text models such as MiniLM, a 1-dimensional positional encoding is added where sinusoidal waves of various frequencies are used:

$$\mathrm{PE}(p, 2i) = \sin(p/10000^{2i/d}) \quad (1)$$

$$\mathrm{PE}(p, 2i+1) = \sin(p/10000^{2i/d}) \quad (2)$$

where p represents the current position of word in a sequence and L is the total length of sentence and i has a range of $0 <= i < L/2$ and the input embedding dimension is labelled as d. The dimensions of input embedding as well as of positional encoding are kept same so they can be added. Image Transformer models such as ViT also used the same positional encoding as using 2-dimensional positional encoding, no significant performance was observed.

B. ENCODER

A stack of N identical layers stacked together. Each layer is composed of two sub-layers. Multi-headed Self Attention is first sub-layer followed by a simple positionally linked feed-forward network. In order to rtain previous information and avoid vanishing gradient problem a residual connection is employed along with layer normalization, after each sub-layers. The LayerNorm of each sub-output layer is equal to LayerNorm(x + Sublayer(x)), where Sublayer(x) is the function that each sub-layer implements on its own. The dimension of all sub-layers, input, and output of the model is kept the same for assisting residual connections.

1) IMAGE INPUT REPRESENTATION

The encoder accepts an image as input and downsample or upsample it to a predetermined size (H, W). While the adjusted picture's size are guaranteed to be divisible by the patch size P, the encoder separates the input image into a batch of patches with defined sizes.The Transformer encoder, by nature, cannot process a whole entity until there are a series of input tokens or in case of images there must be series of patches. The patches are then linearly projected onto D-dimension vectors, where D is the hidden size of the Transformer across all of its layers, after being flattened into vectors.

"[CLS]", a unique token, typically used for the image classification task is retained, much like ViT and DeiT. The entire image is represented by the "[CLS]" token, which combines the data from every patch embedding. The special distillation token employed in DeiT is also kept when utilising the DeiT pre-trained models for encoder initialization so that the student model can distill knowledge from the teacher model. Given learn-able 1D position embeddings based on their absolute positions are the patch embeddings and two special tokens. Afterward, the series of input is fed to a stack of identical encoder layers. MSA and feed-forward fully connected sublayer are present in each Transformer layer. Layer normalisation and residual connection come after these.

2) SELF ATTENTION

The attention mechanism involves dividing up the attention among the values and producing the weighted sum of them, where the weights of the values are determined by the relevant keys and queries. All of the queries, keys, and values for the self-attention modules originate from the same sequence. The attention output matrix is calculated as follows:

$$\text{Self Attention}\ (Q, K, V) = \mathrm{softmax}((Q \cdot K^{\mathrm{T}})/\sqrt{d_k}) \cdot V \quad (3)$$

The softmax function's extremely small gradients are avoided by applying the scaling factor $\sqrt{d_k}$, where dimension of query and key vector is represented by d_k. The model may jointly gather data from several representation subspaces thanks to the MSA sub-layer, which projects the queries, keys, and values h times with various learnable projection weights.

$$\mathrm{Multihead}(Q, K, V) = \mathrm{concat}(h_1 \ldots h_n) \cdot W^o \quad (4)$$

where

$$\mathrm{head}_i = \text{Self Attention}(W_i^q q_j, W_i^k K, W_i^v V)$$

Figure 2 illustrates the calculation of a multi-headed self attention layer using heat maps. A vanilla image transformer model such as ViT have the same architecture of an encoder of [7]. The only difference is that Image transformer takes image as series of input and [7] takes text in a sequence. The self attention calculation mechanism is also same.

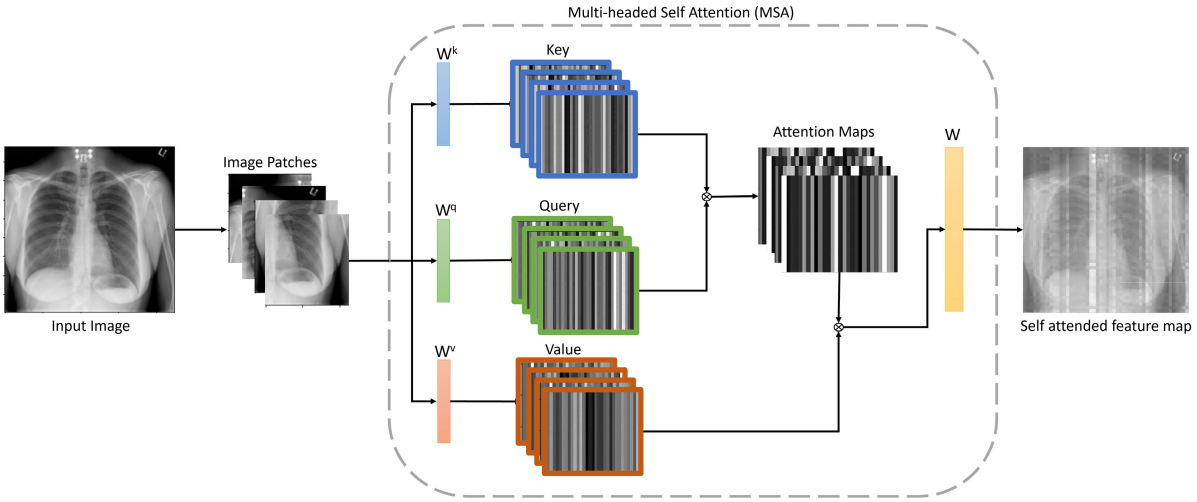


FIGURE 2. Graphical representation of Multi-headed self-attention calculation in a single layer of Transformers.

C. DECODER

Decoder also contains N number of identical layers stacked such as in Encoder but with additional sub-layer. The decoder performs multi-head attention over the hidden states from Encoder, which is then transformed into Queries and Keys, hence inserting a third sub-layer along with two sub-layers already present in each Decoder. Each of the sub-layers applies residual connections in a manner similar to the encoder before layer normalisation. Decoder uses ''Masked Multi-head attention'' layer in order to stop paying attention to succeeding locations. The prediction of location i can only be based on the known outputs at positions lower than i which is due to the masking technique along with the shifting of output embedding offset by one place because of Casual Language Modelling (CLM).

1) TEXT INPUT REPRESENTATION

A sequence of tokens is given as input to decoder. Using the Embedding look-up table, of V vocabulary length, the feature vectors of d_{model} dimension of each token is fetched and positional encoding of sequence is added. At time of training whole report is given as decoder input but at inference only initial start token is given to produce next words using CLM.

2) SELF ATTENTION

For TrMRG, we employ the original Transformer decoder. With the exception of inserting ''encoder-decoder attention'' between the multi-head self-attention and feedforward network to distribute various levels of attention on the encoder output. Similar to Encoder layer, the Decoder also has a stack of identical layers. The keys, values and the queries in the encoder-decoder attention module are derived from the encoder output and the decoder input, respectively. In order to avoid receiving more information during training than necessary for prediction, the decoder also makes use of attention masking in self-attention. In literature, the ''Encoder-Decoder attention'' is also stated as cross attention.

TABLE 3. Comparison of different Image Transformers.

Sr#	Parameters	Encoder	Decoder
1	Model	Vit	MiniLM
2	Hidden layers	24	12
3	Attention heads	16	12
4	Intermediate size	4096	1536
5	Hidden size	1024	384
6	Patch size	32	-
7	Number of patches	144	-
8	Image size	384	-
9	Vocabulary size	-	30522
10	Max length	-	173
11	Beams	-	4
12	Layer normalization	$1e^{-12}$	$1e^{-12}$

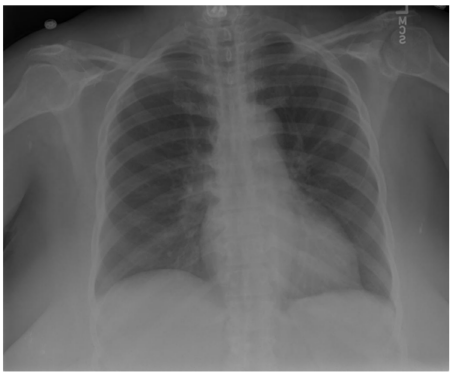
IV. EXPERIMENTS & RESULTS

The proposed model has been tested using chest X-rays images and associated reports. The details are given in following sections

A. DATASET

The Indiana University Chest x-Ray dataset (IU X-Ray) contains a collection of pairs of Chest x-ray images and diagnostic of reports [31]. The dataset includes 3955 reports produced by radiologists and 7,470 pairs of frontal and lateral chest x-ray pictures. Each report consists of different sections e.g. impression, findings, tags, comparison, and indication. In this research, we solely use frontal CXR as input and treat the contents of findings as the target captions to be generated. Figure 4 gives an illustration of a randomly selected frontal view along with associated findings.

The dataset, specially images, are pre-processed to support the implementation of proposed model. Reports (Text data) is cleaned by removing unnecessary spaces, special characters



Impressions:

1. No evidence of acute cardiopulmonary process. 2. 8mm nodule in the left lung base, XXXX calcified granuloma

Findings:

The cardiac and mediastinal contours are within normal limits. The lungs are well-inflated and clear. There is an 8mm nodule in the left lower lobe, XXXX calcified granuloma. There is no pneumothorax or effusion. Bony structures of the thorax are intact with minimal early degenerative change.

FIGURE 3. Data sample containing CXR, Impressions and Findings from IU dataset.

including comma and full stop and hidden words such as "xxxx", also all characters are converted into lower case. A total of 3203 reports are selected for our research and 2404 random samples are selected for Training set, 500 for Validation set and 300 for Testing set.

B. IMPLEMENTATION DETAILS

The pre-trained weights for the encoders and decoders are utilised to fine-tune them on the IU X-Ray dataset during the training phase. The proposed model is implemented using PyTorch and trained on a GeForce RTX 2070 8GB GPU. The AdamW optimizer is used to fine-tune the network for 25 epochs with a batch size of 1. For testing, the parameter values that produce the best results on the validation dataset are employed.

C. EVALUATION METRICS

To evaluate our results, we adopt BLEU [32], METEOR [33], ROUGE-L [34] and CIDEr [35] metrics that are widely-used to evaluate the natural language generation model. The evaluation of machine translation is originally intended to focus on BLEU and METEOR in particular. ROUGE-L is a tool for evaluating summary quality. CIDEr is made to assess image captioning applications.

D. ABLATION STUDY

To evaluate the proposed report generation model, we have conducted detailed ablation studies. All of our experiments are on IU dataset and we reported both qualitative and quantitative results.

1) QUANTITATIVE RESULTS

Our TrMRG model consists of pre-trained image and language model on natural images or text datasets so in first ablation study, we experimented with 4 image models as encoder and 3 Language models as decoder. The qualitative outcomes of ablation study are shown in Table 4. The n-gram similarity between the generated sentences and the ground-truth phrases is the foundation for the automatic evaluation measures (like BLEU). The table shows that MiniLM based decoder in general performed well while generating radiology reports. For encoder, Dino, Deit and ViT produced better results respectively.

TABLE 4. Comparison of different Image Transformers.

Sr#	Encoder	Decoder	B1	B2	B3	B4
1	ViT	MiniLM	0.532	0.344	0.2330	0.158
2	Deit		0.5335	0.3446	0.2294	0.1254
3	Dino		**0.5551**	0.3533	0.2345	0.1332
4	Vit_MAE		0.4339	0.2465	0.1233	0.0513
5	ViT	GPT2	0.4261	0.2441	0.1224	0.053
6	Deit		0.4350	0.2507	0.1264	0.0516
7	Dino		0.5023	0.3312	0.2291	0.1310
8	Vit_MAE		0.4280	0.2431	0.1218	0.0513
9	ViT	DistilGPT2	0.4412	0.2623	0.1478	0.0591
10	Deit		0.4569	0.3038	0.2104	0.1013
11	Dino		0.5349	0.3763	0.2733	0.1726
12	Vit_MAE		0.5405	**0.3804**	**0.2787**	**0.1770**

2) QUALITATIVE ANALYSIS

The IU dataset is quite imbalanced and shows biasness due to more normal reports. So instead of just relying on quantitative values as presented in table 4, we also performed qualitative ablation study. In this section, we evaluate the overall quality of generated reports through several examples. Figure 4 presents reports generated by all 12 models from ablation study for two different scenario of radiology reports. It can be observed that in first scenario when example is from the most occurring cases, reports generated by all models are following the ground truth. The reports generated by all models are clear and provide more detail which is closer to reference report. However, when we look at second scenario which is one of the rare occurring scenario, the models behaviour is different. Most of the models generate first scenario reports due to biased dataset. The ablation study shows that ViT alongwith MiniLM produces the best qualitative results that is why we have selected this combination in proposed model. In general, our model can produce descriptions that follow the logical progression of radiologists' reports, which begin with broad information such views, positive findings, followed by

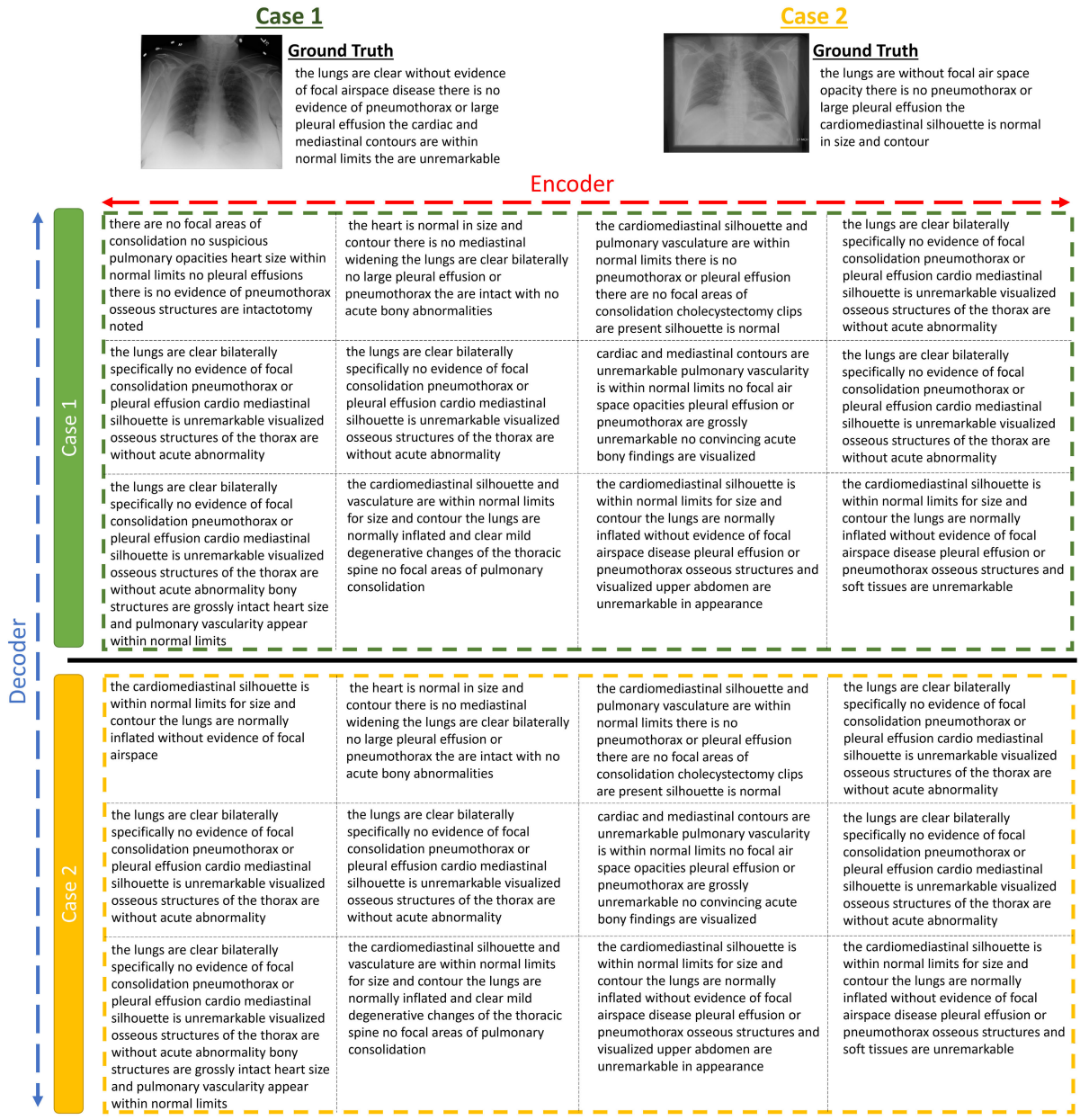


FIGURE 4. Illustration of reports generated by all 12 models for 2 randomly selected cases from IU dataset. From left to right results of ViT, DEIT, DINO and Vit_MAE models are shown and from top to bottom all decoder models MiniLM, GPT2 and DistilGPT2 results are shown respectively.

negative discoveries, in the sequence of lung, heart, pleura, and others.

3) COMPARISON WITH LITERATURE

After finalizing the model through ablation study, we also perform comparative analysis with existing state of the art techniques. Table 5 shows results on the automatic metrics for the Findings module compared to literature. CDGPT2 used Transformer with pre-trained weights at Decoder only whereas for Encoder a Traditional CNN was utilized pre-trained on CXR images and yet still managed to score 0.38 BLEU-1 score. In our TrMRG model, we have employed Transformer based both Encoder and Decoder pre-trained on natural images and language and scored the highest BLEU-1 score. Other researches for instance Align Transformer and Auto Encoder Transformer have made use of Transformer Network at Encoder and Decoder but the selection of ecnoder and decoder caused lower BLEU-1 for these. In comparison with all these techniques on IU dataset, TrMRG model has shown highest BLEU-1 score which indicates its overall efficiency for generating reports that resemble those written by radiologist.

TABLE 5. Comparison with review literature.

Sr#	Model	BLEU 1	BLEU 2	BLEU 3	BLEU 4	ROUGE	CIDER	METEOR
1	RTMIC [13]	0.350	0.234	0.143	0.096	-	0.32	-
2	KERP [14]	0.482	0.325	0.226	0.162	0.339	0.280	-
3	BASE+RM+MCLN [15]	0.470	0.304	0.219	0.16	0.371	-	-
4	RGNet [16]	0.464	0.301	0.212	0.158	-	-	-
5	KGAE-Supervised [17]	0.512	0.327	0.240	0.179	0.383	-	-
6	MV+T+I [18]	0.495	**0.360**	0.278	0.224	0.390	-	-
7	PPKED [19]	0.483	0.315	0.224	0.168	0.376	0.351	-
8	M^2triprogressive [20]	0.486	0.317	0.232	0.173	0.390	-	-
9	CDGPT2 [21]	0.387	0.245	0.166	0.111	0.289	0.257	0.164
10	AlignTransformer [22]	0.484	0.313	0.225	0.173	0.379	-	-
11	Auto Encoder + Transformer [23]	0.497	0.357	**0.279**	**0.225**	**0.408**	-	-
12	Ours	**0.532**	0.344	0.233	0.158	0.387	**0.50**	**0.218**

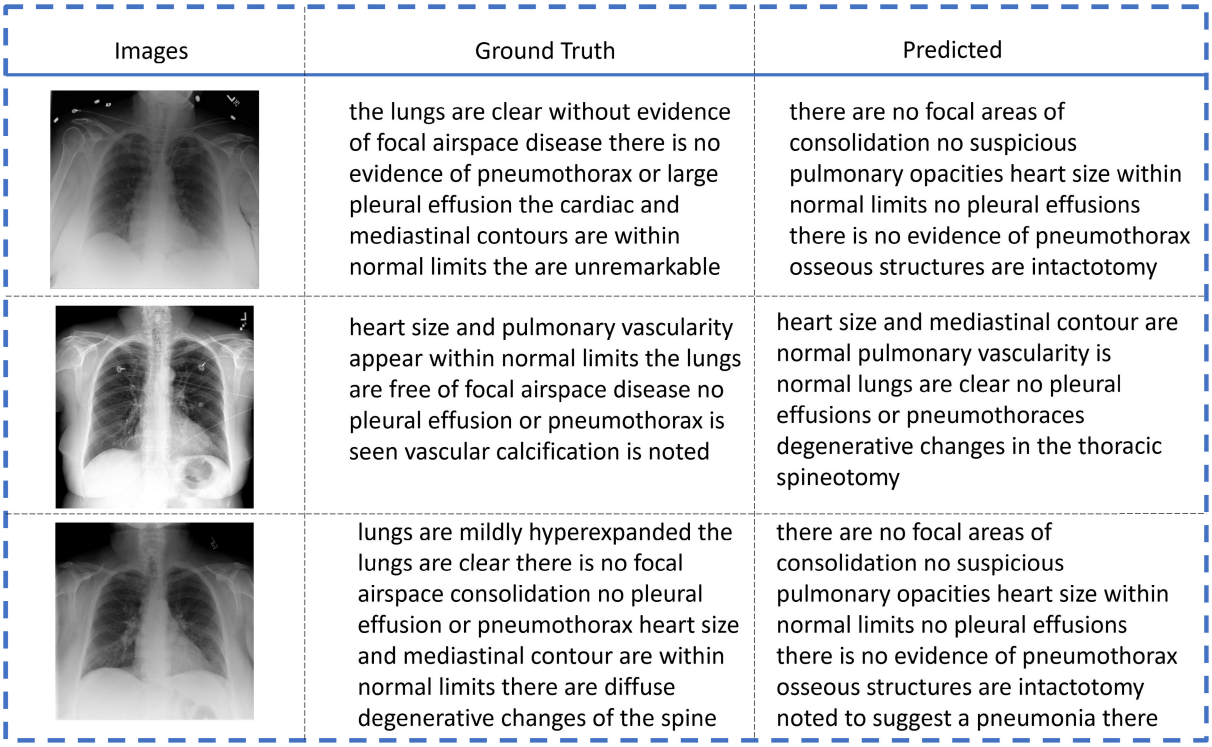


FIGURE 5. Illustration of reports generated by TrMRG. The left most columns shows the CXR of patient given as input and the middle column represents the ground truth whereas the last column shows the generated report from TrMRG.

E. DISCUSSION

Transformer models are still in evolutionary phase for learning how to generate complete paragraphs of medical reports from sparse dataset. Particularly, it is noted that the majority of reports are composed of sentences that are repetitive and strikingly identical, which are of a descriptive nature and do not explain anomalies and disorders. The doctor's reports frequency plot of distinct sentences reveals a long tail of distribution, with anomalous sentences frequently appearing with a frequency of $f = 1$ across the whole dataset. In fact, $f < 3$ occurs in 6,290 of the 8,022 different sentences [36]. Figure 5 exemplify the results of our selected model TrMRG on 3 different cases. It can be noticed that diagnoses of our model relatively similar to doctors report. However at example 3, we can see that our model missed to highlight the abnormality in CXR and this is all due to lack of abnormal samples.

The results show that proposed TrMRG model has performed well when compared to existing models especially in terms of BLEU-1 score. The experimental analysis shows that BLEU score is bit biased towards unbalanced dataset. In case of medical reports where most of the data consists of normal scenarios and reports, a model can quickly achieve good ratings on these automatic evaluation metrics [4] by producing just normal findings. However in proposed model, the learning is bit generic and it is able to generate different reports based on varying inputs. In order to overcome the limitation of BLEU scores, we have also evaluated the proposed model in form of METEOR, ROUGE-L and CIDEr. Table-5 showed that the proposed model has performed

better in majority of these matrices. However, from figure 5, it is clear that for all normal cases the reports generated by proposed TrMRG is similar to reference report but in cases of diseases the model tends to miss some important medical terms. The main reason is quite small corpus size and unique medical terms in IU dataset. This is one of the main limitation of proposed model right now which we intend to overcome with help of large medical corpus along with modification in the model to further refine it for less occurring words.

V. CONCLUSION

In this paper, we proposed a novel approach (TrMRG) to generate radiology reports without using convolutional neural networks which inherits spatial biasness. TrMRG used Image based Transformer pre-trained on Imagenet dataset and Language models pre-trained on natural language datasets. Both encoder and decoder were fine-tuned on IU dataset. The experiments proved effectiveness of our method, which not only generated meaningful reports, but also achieved competitive BLEU score as compared to other models. It is observed that even transformer being data hungry in nature and having no inductive biasness in it can be utilized if properly fine-tuned on smaller datasets such as IU dataset. The ablation study in form of qualitative and quantitative results justified the usefulness of TrMRG in assisting the radiologist.

REFERENCES

[1] O. Er, N. Yumusak, and F. Temurtas, "Chest diseases diagnosis using artificial neural networks," *Expert Syst. Appl.*, vol. 37, no. 12, pp. 7648–7655, Dec. 2010.

[2] T. Gupte, A. Knack, and J. D. Cramer, "Mortality from aspiration pneumonia: Incidence, trends, and risk factors," *Dysphagia*, vol. 37, no. 6, pp. 1493–1500, Dec. 2022.

[3] L. Delrue, R. Gosselin, B. Ilsen, A. Van Landeghem, J. de Mey, and P. Duyck, "Difficulties in the interpretation of chest radiography," in *Comparative Interpretation of CT and Standard Radiography of the Chest*. Berlin, Germany: Springer, 2011, pp. 27–49.

[4] B. Jing, P. Xie, and E. Xing, "On the automatic generation of medical imaging reports," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*. Melbourne, VIC, Australia: Association for Computational Linguistics, Jul. 2018, pp. 2577–2586. [Online]. Available: https://aclanthology.org/P18-1240

[5] I. Allaouzi, M. B. Ahmed, B. Benamrou, and M. Ouardouz, "Automatic caption generation for medical images," in *Proc. 3rd Int. Conf. Smart City Appl.* New York, NY, USA: Association for Computing Machinery, Oct. 2018, pp. 1–6.

[6] J. Yuan, H. Liao, R. Luo, and J. Luo, "Automatic radiology report generation based on multi-view image fusion and medical concept enrichment," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019* (Lecture Notes in Computer Science), vol. 11769. Midtown Manhattan, NY USA: Springer, 2019, pp. 721–729.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. U. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30. Red Hook, NY, USA: Curran Associates, 2017, pp. 1–11.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1. Minneapolis, MN, USA: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[9] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," OpenAI, San Francisco, CA, USA, Tech. Rep., 2019.

[10] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020.

[11] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," in *Proc. NeurIPS EMC Workshop*, 2019, pp. 673–680.

[12] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers," in *Proc. Adv. Neural Inf. Process. Syst., Annu. Conf. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, Dec. 2020.

[13] Y. Xiong, B. Du, and P. Yan, "Reinforced transformer for medical image captioning," in *Proc. Int. Workshop Mach. Learn. Med. Imag.* Midtown Manhattan, NY, USA: Springer, 2019, pp. 673–680.

[14] C. Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Knowledge-driven encode, retrieve, paraphrase for medical image report generation," in *Proc. 33rd AAAI Conf. Artif. Intell. 31st Innov. Appl. Artif. Intell. Conf. 9th AAAI Symp. Educ. Adv. Artif. Intell. (AAAI/IAAI/EAAI)*. Palo Alto, CA, USA: AAAI Press, 2019, pp. 6666–6673, doi: 10.1609/aaai.v33i01.33016666.

[15] Z. Chen, Y. Song, T.-H. Chang, and X. Wan, "Generating radiology reports via memory-driven transformer," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*. Stroudsburg, PA, USA: Association for Computational Linguistics, Nov. 2020, pp. 1439–1449. [Online]. Available: https://aclanthology.org/2020.emnlp-main.112

[16] P. Srinivasan, D. Thapar, A. Bhavsar, and A. Nigam, "Hierarchical X-ray report generation via pathology tags and multi head attention," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, Nov. 2020, pp. 1–17.

[17] F. Liu, C. You, X. Wu, S. Ge, S. Wang, and X. Sun, "Auto-encoding knowledge graph for unsupervised medical report generation," in *Proc. Adv. Neural Inf. Process. Syst.*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021, pp. 16266–16279. [Online]. Available: https://openreview.net/forum?id=nIL7Q-p7-Sh

[18] H. Nguyen, D. Nie, T. Badamdorj, Y. Liu, Y. Zhu, J. Truong, and L. Cheng, "Automated generation of accurate & fluent medical X-ray reports," in *Proc. Conf. Empirical Methods Natural Lang. Process.* Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 3552–3569.

[19] F. Liu, X. Wu, S. Ge, W. Fan, and Y. Zou, "Exploring and distilling posterior and prior knowledge for radiology report generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13748–13757.

[20] F. Nooralahzadeh, N. P. Gonzalez, T. Frauenfelder, K. Fujimoto, and M. Krauthammer, "Progressive transformer-based generation of radiology reports," in *Proc. Findings Assoc. Comput. Linguistics (EMNLP)*. Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 2824–2832. [Online]. Available: https://aclanthology.org/2021.findings-emnlp.241

[21] O. Alfarghaly, R. Khaled, A. Elkorany, M. Helal, and A. Fahmy, "Automated radiology report generation using conditioned transformers," *Informat. Med. Unlocked*, vol. 24, 2021, Art. no. 100557.

[22] D. You, F. Liu, S. Ge, X. Xie, J. Zhang, and X. Wu, "AlignTransformer: Hierarchical alignment of visual regions and disease tags for medical report generation," in *Proc. 24th Int. Conf., Med. Image Comput. Comput. Assist. Intervent. (MICCAI)*, Strasbourg, France, Sep./Oct. 2021.

[23] Y. Wang, Z. Lin, Z. Xu, H. Dong, J. Tian, J. Luo, Z. Shi, Y. Zhang, J. Fan, and Z. He, "Trust it or not: Confidence-guided automatic radiology report generation," 2021, *arXiv:2106.10887*.

[24] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran, "Image transformer," in *Proc. 35th Int. Conf. Mach. Learn.*, vol. 80, J. Dy and A. Krause, Eds., Jul. 2018, pp. 4055–4064. [Online]. Available: https://proceedings.mlr.press/v80/parmar18a.html

[25] A. Kolesnikov, A. Dosovitskiy, D. Weissenborn, G. Heigold, J. Uszkoreit, L. Beyer, M. Minderer, M. Dehghani, N. Houlsby, S. Gelly, T. Unterthiner, and X. Zhai, "An image is worth 16 × 16 words: Transformers for image recognition at scale," in *Proc. 9th Int. Conf. Learn. Represent. (ICLR)*, 2021.

[26] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, Jul. 2021, pp. 10347–10357.

[27] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9650–9660.

[28] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16000–16009.
[29] M. Sirshar, M. F. K. Paracha, M. U. Akram, N. S. Alghamdi, S. Z. Y. Zaidi, and T. Fatima, "Attention based automated radiology report generation using CNN and LSTM," *PLoS ONE*, vol. 17, no. 1, Jan. 2022, Art. no. e0262209.
[30] M. Li, T. Lv, L. Cui, Y. Lu, D. Florencio, C. Zhang, Z. Li, and F. Wei, "TrOCR: Transformer-based optical character recognition with pre-trained models," in *Proc. 37th AAAI Conf. Artif. Intell., Assoc. Advancement Artif. Intell.*, Washington, DC, USA, 2021.
[31] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald, "Preparing a collection of radiology examinations for distribution and retrieval," *J. Amer. Med. Inform. Assoc.*, vol. 23, no. 2, pp. 304–310, Mar. 2016.
[32] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*. Philadelphia, PA, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318. [Online]. Available: https://aclanthology.org/P02-1040
[33] A. Lavie and A. Agarwal, "Meteor: An automatic metric for MT evaluation with high levels of correlation with human judgments," in *Proc. 2nd Workshop Stat. Mach. Transl.*, Jul. 2007, pp. 228–231.
[34] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Proc. Conf. Question Answering Restricted Domains. Assoc. Comput. Linguistics*, Barcelona, Spain, Jan. 2004, p. 10.
[35] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4566–4575.
[36] P. Harzig, Y.-Y. Chen, F. Chen, and R. Lienhart, "Addressing data bias problems for chest X-ray image report generation," 2019, *arXiv:1908.02123*.

MASHOOD MOHAMMAD MOHSAN received the B.S. degree in computer science from the University of Agriculture (UAF), Faisalabad, in 2020. He is currently pursuing the M.S. degree in computer software engineering with the College of Electrical and Mechanical Engineering, National University of Sciences and Technology (NUST), Rawalpindi and doing collaborative research in medical imaging.

MUHAMMAD USMAN AKRAM (Senior Member, IEEE) received the B.S. degree (Hons.) in computer system engineering and the master's and Ph.D. degrees in computer engineering from the College of Electrical and Mechanical Engineering, National University of Sciences and Technology (NUST), Rawalpindi, Pakistan, in 2008, 2010, and 2012, respectively. He is currently an Associate Professor with the College of Electrical and Mechanical Engineering, NUST. He has over 200 international publications in well reputed journals and conferences. His main areas of research interests include biomedical imaging and image processing.

GHULAM RASOOL (Senior Member, IEEE) received the Ph.D. degree in systems engineering from the University of Arkansas, Little Rock, in 2014. He was a Postdoctoral Fellow at the Rehabilitation Institute of Chicago and Northwestern University, from 2014 to 2016. He is currently an Assistant Member with the Machine Learning Department, Moffitt Cancer Center, Tampa, FL, USA. His current research interests include improving cancer care using machine learning, artificial intelligence, and statistical signal and image processing.

NORAH SALEH ALGHAMDI received the Bachelor of Computer Science degree from Taif University, Taif, Saudi Arabia, and the Master of Computer Science and Ph.D. degrees from the Department of Computer Science, La Trobe University, Melbourne, Australia. She is currently an Associate Professor with the Department of Computer Science, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University (PNU), Riyadh, Saudi Arabia. She has been the Vice-Dean of quality assurance, since 2019. She is currently the Director of businesses and projects management in her college. She has authored or coauthored many articles published in a well-known journals in the research field. Her research interests include data mining, machine learning, text analytics, image classification, bioengineering, and deep learning. She has participated in organizing the international conference on computing (ICC 2019). She is a member of the reviewer committee of several journals, such as IEEE Access, *Journal of Computer Science*, and *International Journal of Web Information Systems*.

MUHAMMAD ABDULLAH AAMER BAQAI is currently pursuing the B.S. degree in computer science engineering with Michigan State University (MSU), East Lansing, MI, USA. He is involved in collaborative research in the field of machine learning and medical imaging.

MUHAMMAD ABBAS received the Ph.D. degree from the University of Manchester, U.K. His research interests include software engineering, ERP systems, and machine learning.