Hierarchical X-Ray Report Generation via Pathology tags and Multi Head Attention

Preethi Srinivasan*, Daksh Thapar*, Arnav Bhavsar, and Aditya Nigam

Indian Institute of Technology Mandi, India
{s18001,d18033}@students.iitmandi.ac.in,{arnav,aditya}@iitmandi.ac.in

Abstract. Examining radiology images, such as X-Ray images as accurately as possible, forms a crucial step in providing the best healthcare facilities. However, this requires high expertise and clinical experience. Even for experienced radiologists, this is a time-consuming task. Hence, the automated generation of accurate radiology reports from chest X-Ray images is gaining popularity. Compared to other image captioning tasks where coherence is the key criterion, medical image captioning requires high accuracy in detecting anomalies and extracting information along with coherence. That is, the report must be easy to read and convey medical facts accurately. We propose a deep neural network to achieve this. Given a set of Chest X-Ray images of the patient, the proposed network predicts the medical tags and generates a readable radiology report. For generating the report and tags, the proposed network learns to extract salient features of the image from a deep CNN and generates tag embeddings for each patient's X-Ray images. We use transformers for learning self and cross attention. We encode the image and tag features with self-attention to get a finer representation. Use both the above features in cross attention with the input sequence to generate the report's Findings. Then, cross attention is applied between the generated Findings and the input sequence to generate the report's Impressions. We use a publicly available dataset to evaluate the proposed network. The performance indicates that we can generate a readable radiology report, with a relatively higher BLEU score over SOTA. The code and trained models are available at `https://medicalcaption.github.io`.

1 Introduction

Understanding radiology images such as X-Rays is essential for diagnosis and treatment of many diseases. Given the amount of skill required for accurately reading such images [1], it is challenging for less-experienced radiologists to write medical reports. Hence in healthcare, writing medical reports from X-Ray images becomes a bottleneck for clinical patient care. To aid radiologists, many researchers are investigating the generation of automatic reports from X-Ray images [2,3] by formulating the problem as image captioning [4]. Although Xray report generation task looks similar to a generic image captioning task, there

* Equal Contribution.

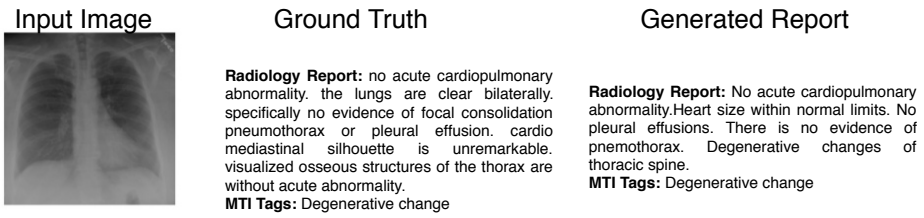


Fig. 1. Shows the actual medical report with MTI tags corresponding to an X-Ray image with the report and tags generated from the proposed network. MTI tags are automatically generated. They are the critical components of the report which capture the essence of the diagnosis.

are fundamental differences and challenges to report generation. The Xray images contain complex spatial information and the abnormalities present in it are difficult to find requiring subject matter expertise. Beyond everything, reports need to be accurate. Hence, we focus on generating clinically accurate reports with reasonably good readability in this work. Figure 1 shows one example of the medical report and tags present in the IU datset [5] with the generated report and tags from our proposed system. Every aspect of the proposed methodology is designed to tackle the challenges present in automatic report generation.

The IU X-ray dataset [5] is used to perform our experiment. Each report in the dataset corresponds to one patient. There is a variable number (N) of X-Ray images of each patient. In the rest of the paper, Pid_{img} refers to a set of N X-Ray images corresponding to a single patient id. Automatically generated tags from the report represent most of the critical components of the report. Findings and Impressions together constitute a report. Tags are identified for each patient, and its embeddings are used in the report generation along with image features. The two parts of the report are generated sequentially. The significant contributions of this paper are as follows:

1. Since in any consortium of diagnostic data a large number of normal patient data exists compared to abnormal patient data, we propose a 2 stage divide-and-conquer approach. First, abnormal patients are identified from normal patients, and their tag embeddings are generated. Conditional learning is done based on the status of the patient's data.
2. For predicting the report, we propose to use a novel architecture involving transformers with 2 Encoders and 2 Decoders instead of traditionally used recurrent neural networks.
3. Tag embeddings and Image features are encoded separately using two Encoders. Findings and Impressions are different and can be learned by two stacked Decoders, helping the former to improve the generation of later.

2 Related Work

An automatic understanding of Radiology images, especially X-ray images, is a well-studied problem. To facilitate that, Wang et al. [6] proposed a large scale dataset for detection and localization of thoracic diseases from X-ray images. They also provided various benchmarks. Yao et al. [7] and Rajpurkar et al. [8] proposed using deep learning-based algorithms for efficient detection of various diseases from chest X-ray images. Later works extended the problem by attributing 'texts' like tags and templates to the x-ray images. Kisilev et al. [9] build a pipeline to predict the attributes of medical images. Shin et al. [10] adopts a CNN-RNN based framework to predict tags (e.g., locations, severities) of chest x-ray images. Zang et al. [11] aimed at generating semi-structured pathology reports, whose contents are restricted to 5 predefined topics.

However, the first work that successfully created an automatic medical report from X-ray images was proposed by Jing et al. [2]. They proposed to use a hierarchical LSTM based recurrent model, exploiting the attention between tags and the image features, opening the field of medical image captioning. Many other works like Wang et al. [3], Li et al. [12], and Xiong et al. [13] enhanced the performance achieved in medical image captioning by proposing various techniques like feature level attention, reinforcement learning, and spatial attention over the localized image regions.

The success in medical image captioning has been possible due to the latest advances in deep learning. DenseNet [14], being a densely connected convolutional network, enabled us to learn high order dependencies by using a large number of layers with a minimal number of parameters, enabling the architectures to understand complex images like X-ray images without overfitting. Xception [15] proposed depth-wise separable convolutional operation, which inturn extracts efficient image features with a decreased number of parameters in the model. Different training strategies like triplet loss function [16] and ranking based loss functions [17,18,19] also enhanced the performance of deep learning based systems for application problems. Moreover, the latest enhances in image captioning problems also played a vital role in developing radiology reports. Karpathy et al. [20] achieved image captioning using deep learning by providing the image features to the initial state of RNN. The RNN then uses the state information to predict the caption of the image. Though RNN's capture temporal dependencies, they have substantial computational overhead. Transformers [21], on the other hand, can efficiently capture long and short term dependencies with minimal computation. Hence, this work tries to utilize the latest deep learning based techniques to generate accurate medical reports of radiology images.

3 Proposed Methodology

This work aims to propose a technique that can generate accurate medical reports using X-ray images of variable sizes (N). Some of the images may cover the neck and abdomen portions too. To avoid the network from getting confused,

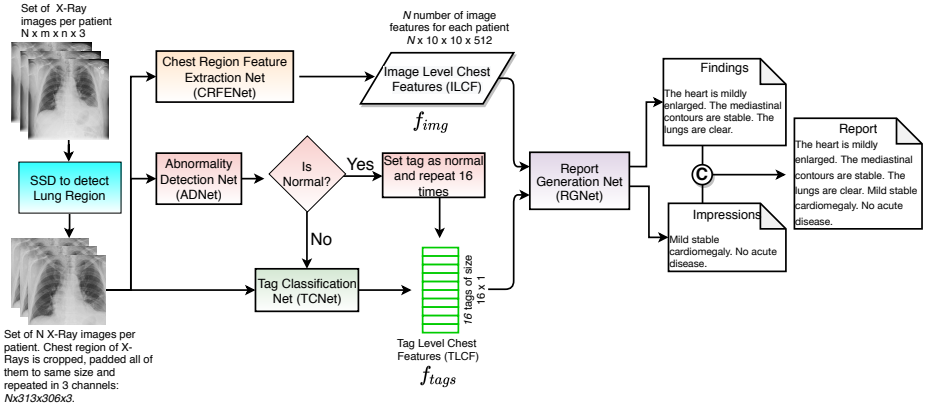


Fig. 2. Shows the overall pipeline of the proposed system. The system's input is a set of X-Rays taken of a patient, and output is the generated medical report containing Findings and Impressions.

first, a Single shot multibox object detector (SSD) [22] is used to detect and crop the lung region from the given X-ray images. The images are then padded to a consistent size of 313×306. For the viability of using pretrained models for extracting image features, we repeat the images in 3 channels of RGB, forming a $313 \times 306 \times 3$ image. Figure 2 shows the overall pipeline that is proposed for generating medical reports from a patient's X-ray images. It consists of 4 modules, namely (i) Chest Region Feature Extraction Net (CRFENet), (ii) Abnormality Detection Net (ADNet), (iii) Tag Classification Net (TCNet), and (iv) Report Generation Net (RGNet). The CRFENet takes the input X-Ray image (I) of size $313 \times 306 \times 3$ and provides a feature of size $10 \times 10 \times 512$. This module is intended to provide contextual information of the image. ADNet also takes an input X-Ray image (I) and does a binary classification to identify any abnormality present. Since there is data imbalance with more data from healthy patients, a hierarchical classification technique is chosen to classify the samples between healthy and unhealthy classes, allowing conditional learning. Only the abnormal samples are put through the TCNet, which ranks the tags to their relevance to the report. The top 16 tags are chosen for each patient. We take only the top 16 tags because the maximum number of tags associated with any patient is 16. In the case of a normal patient, we manually set all the 16 tags to normal. Then, the RGNet takes image features and tags to generate Findings. Then in step 2, it takes Findings to generate Impressions. Finally, Findings and Impressions are concatenated to form the full report. For efficient training and hyper-parameterization, we employ modular training and modular hyper-parameterization. This section discusses each module, the training procedure, and the hyper-parameterization strategy in detail.

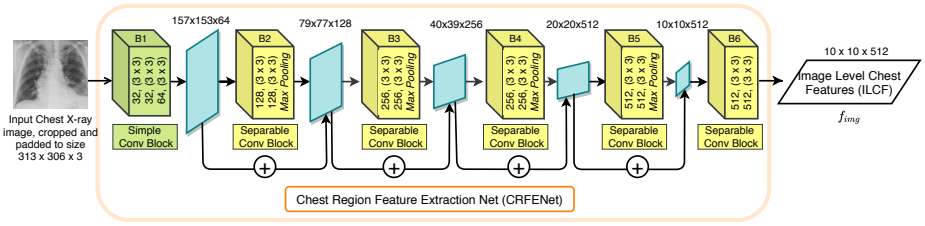


Fig. 3. Shows the architecture of the proposed Chest Region feature extractor. The module contains residual blocks of depth-separable convolutions to decrease the number of overall parameters and computations. It helps to eliminate the over-fitting issues with medical datasets in which the available data is scarce.

3.1 Chest Region Feature Extractor Net (CRFENet)

The first task in generating automatic radiology reports is to identify the salient features present in X-Ray images that lead to the diagnosis. However, the challenge is that these features are complex to recognize and prone to subjectivity; they are highly non-linear. Hence, for extracting such features, we need to learn a complex non-linear function that maps an input image (I) to its feature ($fimg$)) as shown in Figure 3. We have designed a deep convolutional neural network (CNN) for extracting such sophisticated non-linear features. However, using deep CNNs have other disadvantages, such as large number of parameters and vanishing gradient problems. Since medical datasets are scarce (this dataset has only around 3999 patient records), learning deep networks is difficult. We chose to ease the job by incorporating the following two ideas in CRFENet - (i) Use Depth wise separable convolutions [15] over simple convolutions to reduce the number of parameters. (ii) Use residual connections to solve the vanishing gradient problem of deep networks. CRFENet contains one block of simple convolutional layers and four blocks of depth-wise separable convolutional layers, as shown in figure 3. Batch-normalization and Relu non-linearity are used after each conv layer.

Separable Convs: In convolution operation, the kernel aggregates the input feature map's depth information to produce a single output. Hence, to generate an output having depth d, d such kernels are applied, giving us a vast amount of parameters that need to be optimized. Whereas in depth-wise separable convolution operation, one kernel is applied without aggregating depth-wise information. Instead, apply d pointwise convolution kernels to provide us with the final feature with depth d. Using this technique, we can efficiently create a deep model with few parameters avoiding the overfitting problem.

Training: Learning a highly complex non-linear function to map image to its features is a difficult job, especially when the data is scarce. For the CRFENet to understand the X-Ray images, we trained it for chest disease classification on NIH Dataset [6]. The image features of IU Dataset extracted from CRFENet for final report generation are found to be better than the features extracted from deep CNNs like Dense121 [14], VGG [23], and ResNet [24].

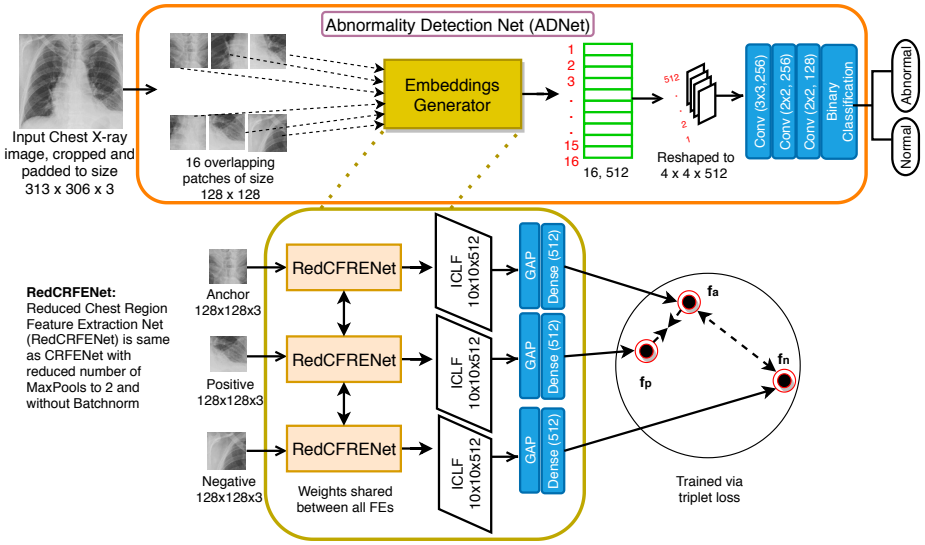


Fig. 4. Shows the architecture of the Abnormality Detection Net (ADNet). It identifies the presence or absence of abnormality in an X-Ray image using the triplet loss function.

3.2 Abnormality Detection Net (ADNet)

As indicated earlier, before we generate medical reports, we need to find whether a patient has any abnormality required to be included in the generated report. To detect abnormal cases, we have proposed a binary classification module called ADNet. ADNet classifies X-ray images into normal and abnormal classes. The patients who do not have any MTI tag associated with them are defined as normal patients. Figure 4 shows the detailed architecture of the proposed ADNet. As the abnormalities present in X-ray images are usually localized, it processes the images in patches. Each input image I is divided into 16 overlapping patches of size 128x128x3. These patches are passed through a sub-network called Embeddings Generator (EG), which produces embedding of size (512x1) for each patch. EG is trained to produce embeddings such that normal patches and abnormal patches are as far as possible from each other in feature space. The 16 embeddings corresponding to a single I are concatenated to form a feature vector of size (16x512). It is further reshaped to (4x4x512) to preserve the spatial relationship present between these patches. Upon applying 2 Convolutional layers and a fully connected layer of 1 neuron to (4x4x512) feature vector, ADNet gives a probability of abnormality. Since every patient's dataset contains a variable number (N) of X-Ray images, we take the average probability and threshold it at 0.5 to classify the patient as normal or abnormal.

Embeddings generator (EG): As discussed above, EG's task is to extract a 512-D feature from a patch (128x128x3) of the X-ray image. EG is trained via triplet loss function [16] to discriminate between normal and abnormal patches.

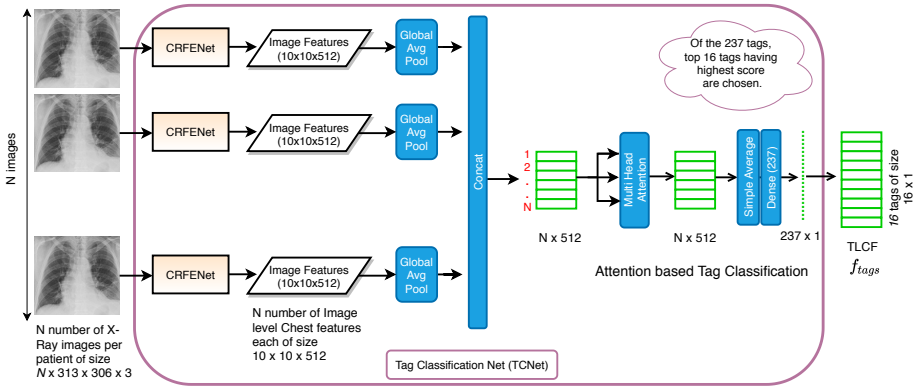


Fig. 5. Shows the architecture of the proposed Tag Classification Net (TCNet). It generates the top 16 relevant tags about a set of X-Ray images of an abnormal patient.

Each patch of size 128x128x3 is passed through rCRFENet, a reduced version of CRFENet, to produce the output of the same size 10x10x512 feature for a 128x128x3 patch. rCRFENet only contains two maxpool layers as compared to 4 in CRFENet. rCRFENet is pretrained on NIH data [6] because it contains the localization information of abnormality in X-Ray images, through an ROI. Patches of 128x128x3 are chosen around the ROI for training. Given two patches i and j, the EG must produce an embedding Θ, such that if both i and j lie in the same class (normal or abnormal), then $L_2(\Theta^i, \Theta^j)$ should tend to 0, otherwise, $L_2(\Theta^i, \Theta^j) \geq \beta$, where β is the margin. The loss has been defined over 3 embeddings:

1. Θ^i: embedding of an anchor patch,
2. Θ^{i^+}: embedding of another patch from the same category, and
3. Θ^{i^-}: embedding of a patch from other categories.

Formally:

$$\mathcal{L}(i, i^+, i^-) = \max(0, (\Theta^i - \Theta^{i^+})^2 - (\Theta^i - \Theta^{i^-})^2 + \beta) \tag{1}$$

We sum the loss for all possible triples (i, i^+, i^-) to form the cost function J which is minimized during training of EG:

$$J = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(i, i^+, i^-) \tag{2}$$

For efficiently training the EG network, we apply online semi-hard negative mining and dynamic adaptive margin as proposed by [25].

3.3 Tag Classification Net (TCNet)

As the second step of hierarchy, TCNet predicts the tags associated with each Pid_{img}. MTI tags play a crucial role in generating the report. As shown in

Figure 5, N images in (Pid_{img}) are passed through CRFENet one after another to obtain image features (10x10x512). Upon applying Global Average Pooling to each of the image features, we get (1x512) feature vector, concatenated to get an (Nx512) feature vector. This Nx512 feature vector passed through a Multi-Head Attention module (MHA) to get the same dimensional output. MHA checks the information in all the N images and produces a result. Over that, a simple averaging and Dense layer of 237 neurons is applied. TCNet is trained using log sum exponential pairwise loss function [19]. It assigns a value to each tag relative to other tags by learning to rank via pairwise comparisons. Then, the values are sorted, and the top 16 tags are picked to produce an output of size (16x1).

Multi Head Attention (MHA): The basic building block of multi-head attention [21] is the scaled dot product mechanism. The scaled dot product mechanism is a sequence to sequence operation: given a sequence of values vectors v_1, v_2, ..., v_n, it learns to provide an output sequence vectors y_1, y_2, ..., y_n, based on a query sequence q_1, q_2, ..., q_n, and key sequence k_1, k_2, ..., k_n. where each vector in the sequence is d-dimensional. First we learn three weight matrices of size $d \times d$ to transform each of the three sequences: $Q_i = W_q q_i \quad K_i = W_k k_i \quad V_i = W_v v_i$. Each y_i is computed as weighted average over the all transformed value vector V: $y_i = \sum_j w_{ij} V_j$, where j iterates over the whole sequence. Here w_{ij} is derived from dot product of query and key sequences: $w'_{ij} = \frac{Q_i^T K_j}{\sqrt{d}}, \quad w_{ij} = softmax(w'_{ij})$. Alternatively, in the scaled dot product mechanism, to compute one particular output y_i, the corresponding vector of query Q_i is compared (via. dot product) to the whole sequence of key vectors K_1, K_2, ..., K_n to provide the attention weights for each of the value vectors V_1, V_2, ..., V_n. We use the scaled dot product mechanism to form the multi-head attention mechanism. For a given set of value, query, and key vectors of $n \times d$, where n is the sequence length, and d is the dimensionality of each vector; we break each vector into r subparts of $n \times \frac{d}{r}$. We apply r different scaled dot product mechanisms, each having independent weight matrices of $\frac{d}{r} \times \frac{d}{r}$ giving us r outputs of $n \times \frac{d}{r}$. We concatenate these outputs to get the final output of shape $n \times d$. Here the total number of parameters is only $\frac{3d^2}{r}$ (3 weight matrices for each of r parts of the input sequence).

3.4 Report Generation Net (RGNet)

This is based on transformer architecture inspired from [21]. RGNet consists of 2 Encoders called Data Encoder (E_D) and Tag Encoder (E_T), and 2 Decoders called Findings Decoder (D_{fin}) and Impressions Decoder (D_{imp}). The network architecture is shown in figure 6.

Data Encoder (E_D): It takes N images of (Pid_{img}), passes each of them through CRFENet to get (10x10x512) for each image. N features are concatenated to form a feature of size (Nx10x10x512) and reshaped to (Nx512x100). Since neither all the N images are equally important nor every 512 features, we try to enhance the appropriate features and images using 2 MHA modules. The first one learns self-attention over each of N images' features providing us

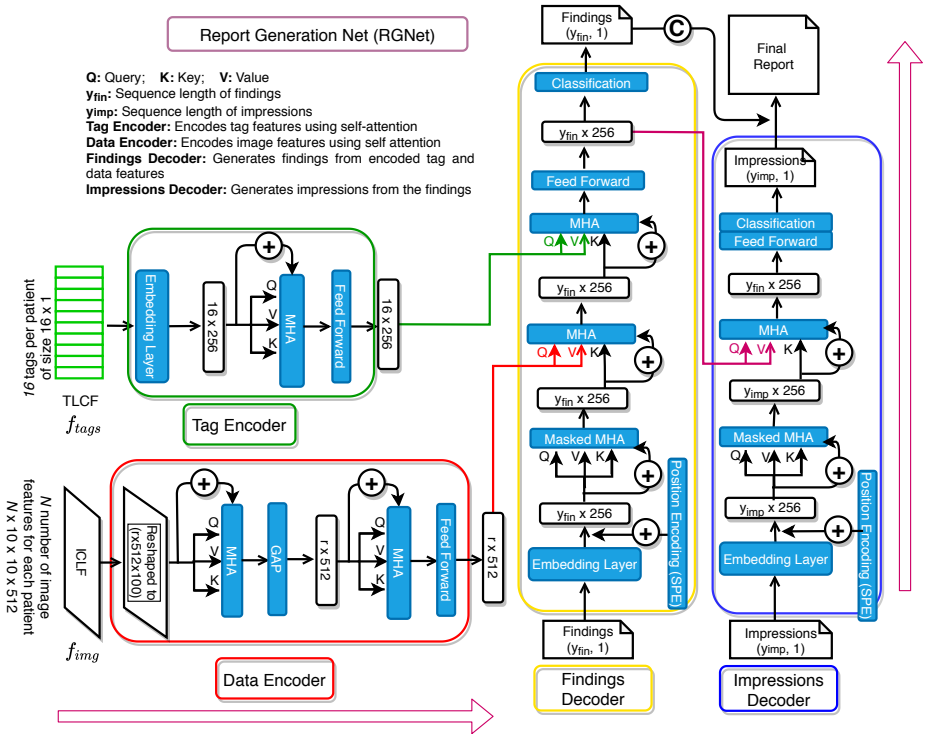


Fig. 6. Shows the architecture of the proposed Report Generation Net (RGNet). This module generates the report using a blend of information from image feature and tag embeddings. Also, sequentially uses the report's Findings to generate the report's Impressions.

with (Nx512) feature map. The second MHA learns self-attention to combine the features across N images forming feature embeddings efficiently.

Tag Encoder (E_T): It takes the (16x1) tags extracted from TCNet and creates an embedding for each of the tags. Later an MHA is used to learn self-attention over the tag embeddings providing us with relevant tags only.

Findings Decoder (D_{fin}): The task of (D_{fin}) is, given a sequence of words corresponding to the Findings, tag embeddings, and image features; it has to generate the next word of the Findings. The next word will depend upon previous words as well as both tag embeddings and image features. We use a transformer block to learn the attention required on previous words of the report over the tags embeddings and image features. Firstly, we learn self-attention on the previous words of the report. It consumes all the previous information to generate the next word. A multi-head attention mechanism is used to learn the self-attention, where the report is given as the key, query, and value. Secondly, we learn cross-attention between the output of the first self-attention block and image features. We again use the multi-head attention mechanism, but for learn-

ing cross-attention over tag embeddings. In both cases, the embeddings are given as value and key, where the previous attention block's output is given as query to multi-head attention block. Self-attention gives us the next word's dependence on previous words, whereas the cross-attention provides us with the dependence of the next word on the image features and tag embeddings. Both Self and Cross attention matrices update their parameters based on the loss generated for next word prediction. A feed-forward layer is applied after the cross attention forming the transformer block. An embedding layer is used to convert the words into embeddings of 256 dimensions, and sinusoidal positional encoding (SPE) [21] is applied over the embeddings before inputting them into the transformer block. Finally, a linear layer followed by softmax cross-entropy loss gives us the probability for each word in the dictionary to be the next word in the Findings.
Impressions Decoder (D_{imp}): Given a sequence of words corresponding to the Impressions and output feature from D_{fin}), (D_{imp}) generates the next word of the Impressions. It first learns the dependence of the next word on previous words by learning self-attention using MHA. Later cross-attention is learned between previous words of Impressions and the generated Findings from D_{fin}) using MHA, enabling the network to produce Impressions depending upon the previously produced Findings. Finally, the Findings and Impressions are concatenated to form the final report.

3.5 Modular Training and Hyper-Parameterization

For training and searching for the optimal hyper-parameters of the proposed methodology, we use modular training. We follow the below sequence of steps, and each model is hyper-parameterized for efficiently performing its pretraining task: (i) Pre-train the CRFENet and rCRFENet for chest disease classification on NIH Dataset [6]. (ii) Then we train the EG using the triplet loss function over patches extracted from the NIH dataset. (iii) Then ADNet is trained over IU-dataset for normal vs. abnormal classification. (iv) The pretrained CRFENet is used to finetune the TCNet using the ranking loss function. (v) Finally, we train the RGNet for report generation using the image features extracted from CRFENet and tags from TCNet.

4 Experimental Analysis

In this section, we provide the details of the experimental analysis performed to validate the proposed methodology. We have performed a thorough ablation study for experimentally validating every contribution proposed in this work. Later, we show that our proposed model can produce accurate reports using qualitative and quantitative comparative analysis.

4.1 Dataset used and evaluation metric

For validating the proposed methodology, we use a publicly available IU X-ray dataset [5]. It contains the medical data of 3999 patients. Each data contains

Task	Model	# layers	# parameters	Performance
Feature Extractor	DenseNet	100	27.2 M	0.175 Loss
	CRFENet	**17**	**12.3 M**	**0.19 Loss**
Abnormality Detection	DenseNet	100	27.2 M	70.5% Acc
	ADNet	**18**	**13.1 M**	**74% Acc**
Tag Classification	TCNet with wBCE	19	12.6 M	0.44 Loss
	TCNet	**19**	**12.6 M**	**0.26 Loss**
Report Generation	**RGNet**	**12**	**0.7 M**	**0.464 Bleu-1**

Table 1. Parametric comparison and modular ablation analysis. Acc: Accuracy, wBCE: weighted Binary Cross Entropy, M: Million. Bold represents the proposed systems.

findings, impressions, MTI tags, and a set of N number of X-ray images taken for each patient. Findings and impressions are combined to make the medical report of the patient. Medical text indexer (MTI) is used to extract keywords from the report forming the MTI tags (referred to as tags in this work). Since each patient has had multiple X-rays, there are a total of 7470 x-ray images. We tokenize each word of the report and remove nonalphabetic tokens. Moreover, we computed the frequency percentile of all the unique words in all the reports and picked only the top 99 percentile of words, which amounts to 1000. The dataset contains 573 unique tags. We take only those tags that appeared in at least three reports. Hence we are left with 283 tags for the tag prediction task. We discarded those patients' data, which did not contain either findings or impressions or X-ray images. For testing the performance of our proposed network, as suggested by Li et al. [12], we randomly split patients for training/validation/testing in the ratio of 7/2/1. For evaluating the report generated against the original report, we use standard image captioning evaluation metric BLEU score (Papineni et al. [26]). BLEU score measures the quality of the text generated and assigns a metric between 0 and 1. It analyses the statistics of overlapping words with the reference sequence. The original report is taken as a reference to run the string matching algorithm. A value of 0 means there is no overlap with the original report, and 1 means there perfect overlap with the original report.

4.2 Ablation Study

For validating the contributions of the proposed network, we have performed an extensive ablation study, as shown in Table 2 and 1. It is important to note, though the system is broken into multiple modules, the complexity of the overall system (38.7 M parameters and 66 layers) is comparable or lesser than state-of-the-art systems. Table 1 shows the parametric comparison and performance of each of the individual modules concerning corresponding state-of-the-art systems. In the first row of the table 2, we have shown our proposed methodology, which contains four modules named CRFENet, ADNet, TCNet, and RGNet, as described in Section 3. We performed an ablation study with alternatives for every network mentioned above to testify each of the proposed networks' contributions. Firstly, we have tested the system by replacing the proposed CRFENet

Model	Bleu-1	Bleu-2	bleu-3	Bleu-4
Proposed Methodology	0.464	0.301	0.212	0.158
Model A(without CRFENet)	0.414	0.287	0.198	0.143
Model B(without ADNet)	0.320	0.218	0.156	0.116
Model C(without ranking loss)	0.295	0.192	0.104	0.092
Model D(without 2 decoder RGNet)	0.423	0.292	0.204	0.148

Table 2. Ablation study of the proposed methodology validating our contributions.

with pre-trained state-of-the-art CNN architectures (Model A). Among such architectures, Densenet provides us with the best performance, as shown in the table's second row. Since CRFENet is only a six-block module with separable convs, fewer parameters prevent overfitting than DenseNet and other CNN modules. Secondly, instead of ADNet, we used a simple VGG network for abnormality detection (Model B). Since most of the X-rays' abnormalities are localized, patch-based siamese abnormality detection (ADNet) provides us better results than standard-sized image-based classifiers like VGG. Thirdly, we trained the TCNet with weighted binary cross-entropy loss rather than ranking loss (Model C). Since most of the tags are only associated with very few reports, the ranking loss is better able to capture the association of tags with particular X-Ray images. Finally, we train a single decoder RGNet, rather than the proposed two decoder RGNet (Model D). The sequentially stacked decoders' training, one for Findings and Impressions, will learn better to optimize their respective models. It will force the network to generate accurate Findings so that better Impressions can get generated and vice-versa. It is evident from table 2 that each of the ablated models performs inferior to the proposed network, which concretely validates every contribution proposed in this work. We can also notice the magnitude of the gain obtained from each of these changes. Usage of ranking loss and hierarchical tag classification techniques gave the relatively biggest deltas in the report generation's quality.

4.3 Comparative Analysis

Table 3 shows the comparative analysis of the proposed system with state-of-the-art networks. It can be seen from the table that our proposed methodology achieves state-of-the-art for report generation task. Key components of our proposed methodology like hierarchical tag classification, ranking based loss, attention-based feature extraction, and transformer architecture could be the leading cause for our model's performance to be better than the rest.

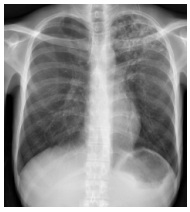
Input Image	Ground Truth	Generated Report
	**Radiology Report:** heart size within normal limits. mild hyperinflation of the lungs. mild pectus excavatum deformity. stable left mid lung calcified granuloma. no focal airspace disease. no pneumothorax or effusions. changes of chronic lung disease with no acute cardiopulmonary finding. **MTI Tags:** Calcified Granuloma	**Radiology Report:** the heart is normal in size. the mediastinum is unremarkable. there is no pleural effusion. pneumothorax. or focal airspace disease. there is stable calcified granuloma in the left lower lobe. no acute disease. **MTI Tags:** Calcified Granuloma
	**Radiology Report:** the heart is mildly enlarged. the mediastinal contours are stable. the lungs are clear. mild stable cardiomegaly. no acute disease. **MTI Tags:** cardiomelagy	**Radiology Report:** the heart is mildly enlarged. the mediastinal contours are stable. there is no pleural effusion. pneumothorax. or focal airspace disease. the lungs are clear. mild stable cardiomegaly. no acute disease. **MTI Tags:** degenerative change, cardiomelagy
	**Radiology Report:** No acute cardiopulmonary abnormality. There are no focal areas of consolidation. No suspicious pulmonary opacities. Heart size within normal limits. No pleural effusions. There is no evidence of pnemothorax. Degenerative changes of thoracic spine. **MTI Tags:** degenerative change.	**Radiology Report:** No acute cardiopulmonary abnormality.Heart size within normal limits. No pleural effusions. There is no evidence of pnemothorax. Degenerative changes of thoracic spine. **MTI Tags:** degenerative change.
Faliure Cases		
	**Radiology Report:** status post midline sternotomy with intact. stable mild cardiomegaly. normal lung vascularity. the lungs are clear. stable postop changes with stable mild cardiomegaly and normal lung vascularity. **MTI Tags:** sterntomy	**Radiology Report:** the heart is normal in size. the mediastinum is unremarkable. mild pectus excavatum deformity is noted. the lungs are clear. no acute disease. **MTI Tags:** pectus excravectum.
	**Radiology Report:** cardiomediastinal silhouette is unchanged with mild cardiomegaly. there is relative elevation of the right hemidiaphragm consistent with history of right lower lobectomy. without focal consolidation. pneumothorax. or effusion identified. irregularity of the right <unk> and <unk> ribs stable since at <unk> and <unk> postsurgical <alt> post traumatic in <unk>. left shoulder rotator <unk> bone <unk> noted. **MTI Tags:** cardiomelagy, lebactomy	**Radiology Report:** No acute cardiopulmonary abnormality. There are no focal areas of consolidation. No suspicious pulmonary opacities. Heart size within normal limits. No pleural effusions. There is no evidence of pnemothorax. Degenerative changes of thoracic spine. **MTI Tags:** degenerative change.

Fig. 7. Shows the qualitative results of report generated from our proposed network. The first 3 rows depict examples from high accuracy outputs. The correctly predicted vocabularies are highlighted. The last 2 rows contain failure cases.

Model	Bleu-1	Bleu-2	bleu-3	Bleu-4
S&T [27]	0.265	0.157	0.105	0.073
SA&T [28]	0.328	0.195	0.123	0.080
TieNet [3]	0.330	0.194	0.124	0.081
Lie et al. [29]	0.359	0.237	0.164	0.113
CNN-RNN [10]	0.216	0.124	0.087	0.066
LRCN [30]	0.223	0.128	0.089	0.067
AdaAtt [31]	0.220	0.127	0.089	0.068
Att2in [32]	0.224	0.129	0.089	0.068
RTMIC [13]	0.350	0.234	0.143	0.096
Li et al. [12]	0.438	0.298	0.208	0.151
CoAtt [2]	0.455	0.288	0.205	0.154
Proposed Methodology	**0.464**	**0.301**	**0.212**	**0.158**

Table 3. Comparative analysis of the proposed system with state-of-the-art.

4.4 Qualitative Analysis

Figure 7 shows the qualitative results of the report generated from our proposed network. The first three rows depict examples from high accuracy outputs, whereas the last two rows contain the failure cases. In the first row, the proposed system can correctly identify calcified granulomas and generate a technically sound report. In the second row, the proposed system identifies cardiomelagy. Moreover, for cases where there were only degenerative changes, our method performed well. We can also understand from the highlighted portion that most of the report's predicted characteristics match the original report. We did find two significant cases of failures; one example of both is depicted in the figure. The first case being those abnormalities that only come ones or twice in the dataset. In such cases, the proposed system was not able to learn about them. The second case of failure is where the images were blurry or hazy. In such cases, the network predicted the patient to have no disease at all.

5 Conclusion

Captioning medical images is a complex task because, unlike the natural images, the salient features are not apparent. Here, we proposed a technique to blend the image and tag features and use it in a unique way to generate a medical report from a patient's set of X-Ray images. Traditional use of recurrent neural networks (RNNs) to solve such sequential data has a massive computational overload. On the other hand, transformer architecture, which also captures the sequential data, uses far fewer parameters. Furthermore, it applies attention between and across features obtained from images, tags, and reports. While significant improvements have been achieved over the SOTA, there is still scope for improvement in generating useful quality reports, especially in hazy X-Rays or cases where different X-Rays are acquired under different exposures.

References

1. Delrue, L., Gosselin, R., Ilsen, B., Van Landeghem, A., de Mey, J., Duyck, P.: Difficulties in the interpretation of chest radiography. In: Comparative interpretation of CT and standard radiography of the chest. Springer (2011) 27–49
2. Jing, B., Xie, P., Xing, E.: On the automatic generation of medical imaging reports. ACL (2018)
3. Wang, X., Peng, Y., Lu, L., Lu, Z., Summers, R.M.: Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 9049–9058
4. Johnson, J., Karpathy, A., Fei-Fei, L.: Densecap: Fully convolutional localization networks for dense captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 4565–4574
5. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association **23** (2016) 304–310
6. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: IEEE CVPR. (2017)
7. Yao, L., Poblenz, E., Dagunts, D., Covington, B., Bernard, D., Lyman, K.: Learning to diagnose from scratch by exploiting dependencies among labels. arXiv preprint arXiv:1710.10501 (2017)
8. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., et al.: Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225 (2017)
9. Kisilev, P., Walach, E., Barkan, E., Ophir, B., Alpert, S., Hashoul, S.Y.: From medical image to automatic medical report generation. IBM Journal of Research and Development **59** (2015) 2–1
10. Shin, H.C., Roberts, K., Lu, L., Demner-Fushman, D., Yao, J., Summers, R.M.: Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 2497–2506
11. Zhang, Z., Xie, Y., Xing, F., McGough, M., Yang, L.: Mdnet: A semantically and visually interpretable medical image diagnosis network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 6428–6436
12. Li, Y., Liang, X., Hu, Z., Xing, E.P.: Hybrid retrieval-generation reinforced agent for medical image report generation. In: Advances in neural information processing systems. (2018) 1530–1540
13. Xiong, Y., Du, B., Yan, P.: Reinforced transformer for medical image captioning. In: International Workshop on Machine Learning in Medical Imaging, Springer (2019) 673–680
14. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 4700–4708
15. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 1251–1258

16. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2015) 815–823
17. Weston, J., Bengio, S., Usunier, N.: Wsabie: Scaling up to large vocabulary image annotation. In: Twenty-Second International Joint Conference on Artificial Intelligence. (2011)
18. Zhang, M.L., Zhou, Z.H.: Multilabel neural networks with applications to functional genomics and text categorization. IEEE transactions on Knowledge and Data Engineering **18** (2006) 1338–1351
19. Li, Y., Song, Y., Luo, J.: Improving pairwise ranking for multi-label image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 3617–3625
20. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2015) 3128–3137
21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. (2017) 5998–6008
22. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European conference on computer vision, Springer (2016) 21–37
23. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
24. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 770–778
25. Thapar, D., Jaswal, G., Nigam, A., Arora, C.: Gait metric learning siamese network exploiting dual of spatio-temporal 3d-cnn intra and lstm based inter gait-cycle-segment features. Pattern Recognition Letters **125** (2019) 646–653
26. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics (2002) 311–318
27. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2015) 3156–3164
28. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning. (2015) 2048–2057
29. Liu, G., Hsu, T.M.H., McDermott, M., Boag, W., Weng, W.H., Szolovits, P., Ghassemi, M.: Clinically accurate chest x-ray report generation. arXiv preprint arXiv:1904.02633 (2019)
30. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2015) 2625–2634
31. Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 375–383

32. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 7008–7024