



Show, tell and summarise: learning to generate and summarise radiology findings from medical images

Sonit Singh¹ · Sarvnaz Karimi² · Kevin Ho-Shon³ · Len Hamey¹

Received: 18 April 2020 / Accepted: 25 March 2021 / Published online: 5 April 2021
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2021

Abstract

Radiology plays a vital role in health care by viewing the human body for diagnosis, monitoring, and treatment of medical problems. In radiology practice, radiologists routinely examine medical images such as chest X-rays and describe their findings in the form of radiology reports. However, this task of reading medical images and summarising its insights is time consuming, tedious, and error-prone, which often represents a bottleneck in the clinical diagnosis process. A computer-aided diagnosis system which can automatically generate radiology reports from medical images can be of great significance in reducing workload, reducing diagnostic errors, speeding up clinical workflow, and helping to alleviate any shortage of radiologists. Existing research in radiology report generation focuses on generating the concatenation of the findings and impression sections. Also, existing work ignores important differences between normal and abnormal radiology reports. The text of normal and abnormal reports differs in style and it is difficult for a single model to learn both the text style and learn to transition from findings to impression. To alleviate these challenges, we propose a *Show, Tell and Summarise* model that first generates findings from chest X-rays and then summarises them to provide impression section. The proposed work generates the findings and impression sections separately, overcoming the limitation of previous research. Also, we use separate models for generating normal and abnormal radiology reports which provide true insight of model's performance. Experimental results on the publicly available *IU-CXR* dataset show the effectiveness of our proposed model. Finally, we highlight limitations in the radiology report generation research and present recommendations for future work.

Keywords Medical imaging · Radiology report generation · Chest X-rays · Computer-aided report generation · Artificial intelligence · Computer vision · Natural language processing · Deep learning

1 Introduction

Artificial intelligence (AI) is one of the most disruptive technologies to a majority of industries, including health care [1]. Medical imaging plays a vital role in health care by viewing the human body for diagnosis, monitoring, and treatment of medical conditions. In radiology practice, radiologists routinely examine medical images and conclude their findings in the form of radiology reports. These textual reports narrate the findings of the abnormalities and diseases present in the medical image. Figure 1 shows a sample chest X-ray with its accompanying radiology report. A radiology report consists of several sections, mainly *comparison*, *indication*, *findings* and *impression*. Examining medical images and writing radiology reports consume most of a radiologists' time. The task of

✉ Sonit Singh
sonit.singh@hdr.mq.edu.au

Sarvnaz Karimi
sarvnaz.karimi@data61.csiro.au

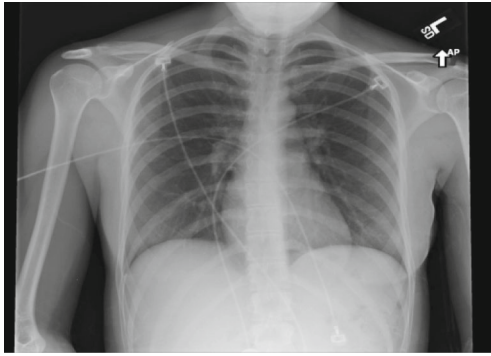
Kevin Ho-Shon
kevin.ho-shon@mq.edu.au

Len Hamey
len.hamey@mq.edu.au

¹ Department of Computing, Macquarie University, Sydney, NSW, Australia

² Data61, CSIRO, Sydney, NSW, Australia

³ Department of Clinical Medicine, Macquarie University, Sydney, NSW, Australia

**Comparison:**

None.

Indication:

Chest pain, feels out of it.

Findings:

The Cardiomediastinal silhouette and pulmonary vasculature are within normal limits in size. The lungs are clear of focal airspace disease, pneumothorax, or pleural effusion. There are no acute bony findings.

Impression:

No acute cardiopulmonary findings.

Fig. 1 An example of a chest X-ray and its corresponding radiology report in the IU-CXR dataset [2]

examining medical images and writing radiology reports is tedious, time consuming, error-prone, often representing a bottleneck in clinical diagnosis process. Also, with improvements in imaging techniques and the wider availability of imaging facilities, there is an increasing demand for imaging studies, resulting in clinical overload. Due to this, numerous scans remain unreported, leading to delayed or missed diagnoses and an adverse impact on patient care.

Computer-aided detection (CADe) and Computer-aided diagnosis (CADx) systems have been an integral part of the Picture Archiving and communication systems (PACS) for medical image analysis. These CAD systems play a vital role in radiology practice by providing a second opinion, reducing labor cost, automating tedious and repetitive tasks, and helping alleviate the situation where the number of experienced medical experts is far below the demand. With the same goal to support radiologists, the task of automatic radiology report generation from medical images is proposed [3–5]. A CAD system which can automatically generate radiology reports from medical images can be great significance in reducing workload, reducing diagnostic errors, speeding up clinical workflow, and help in alleviating the situation where the number of radiologists is lower and medical resources are scarce.

Existing research [5–7] concatenates the findings and impression sections to form the target for radiology report generation task. This method lacks the ability to distinguish the heterogeneous information present in the actual radiology reports. The *findings* section is a detailed description of various pathologies, their location, and their severity. On

the other hand, the *impression* section is a summary of the radiology report and concludes the radiology study. Hence, concatenating findings and impression section into a single text for radiology report generation is not a clear indication of a model's performance, given different nature of the two sections.

Previous research [5–7] trains a single model for radiology report generation from medical images irrespective of whether the images are normal or abnormal. A single model trained with both *normal* and *abnormal* radiology studies biases the model's performance toward generating normal reports given that abnormal studies are rare compared to normal studies. For example, in the IU-CXR data, about 2/3 of radiology reports are normal. Further, training of the single model gives equal weight to both normal and abnormal report studies, ignoring the fact that finding abnormalities and pathologies in medical images is critical in terms of clinical accuracy. Hence, results obtained from a single trained model are not reflective of whether the underlying learning model is able to successfully generate abnormal findings.

Inspired by the fact that radiologists first interpret medical image and write *findings* section to highlight abnormalities present in the image. The findings section is the most detailed description providing in-depth analysis of normal and abnormal regions in an image. It also provide details about location and severity of abnormalities. Based on the findings section along with background context in terms of history, demographics of a patient, radiologists briefly summarise findings into an impression section which concludes the study. On the contrary, existing approaches for radiology report generation concatenate findings and impression section as a text to be generated by the trained model. Given the task of radiology report generation aims to generate reports that are coherent despite also being long, the concatenation of the findings and impression sections into a single text for radiology report generation increases the length of the text to generate, which deteriorates the performance of generation models. In order to mimic radiology workflow and to overcome challenge of modelling long-range dependencies in RNNs, we propose two-tier model separating generation and summarisation parts. The generation part takes in chest X-ray and generates detailed findings and the summarisation part takes in the generated findings to summarise them into an impression section which concludes the study.

To alleviate limitations of the previous work, we propose a *Show, Tell and Summarise* model based on radiology report generation from medical images. A block diagram of our proposed model is shown in Fig. 2. Our proposed model consists of three sequential modules: (1) an *image classification module* which classifies an input chest X-ray as *Normal* or *Abnormal*; (2) a *findings*

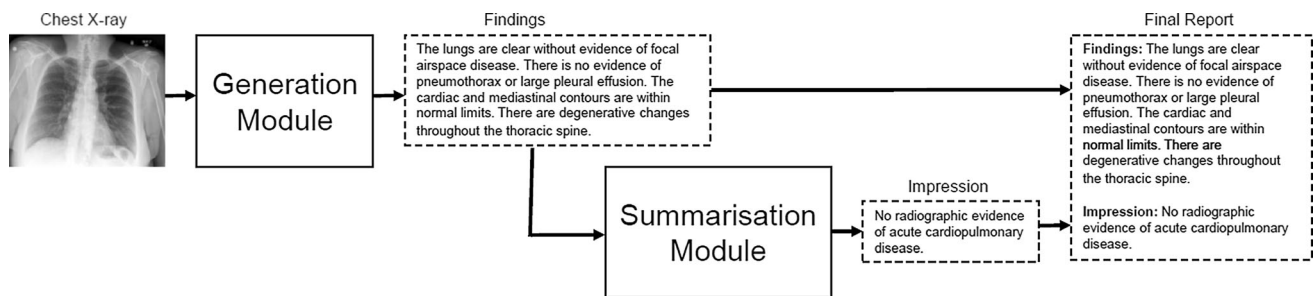


Fig. 2 Block diagram of show-attend-summarise model for generating and summarising radiology findings from medical images

generation module, which is an *encoder–decoder* framework [8] that takes either normal or abnormal chest X-ray as an input and generates *findings* as an output; and, (3) a *summarisation module*, which summarises the generated *findings* and gives an *impression* section as an output. We evaluate the proposed model on the publicly available Indiana University Chest X-ray Collection (IU-CXR) [2] dataset. The experimental results show that the proposed model has capability to generate better findings and impression sections separately, given an input chest X-ray to the model.

The contributions of the paper are:

1. We present a novel architecture which can simultaneously classify a medical image, generate findings for a given medical image, and summarise the generated findings to form an impression of a radiology report;
2. We propose to model normal and abnormal reports separately, given that the linguistic characteristics of normal and abnormal reports are quite different in terms of syntactic and semantic information; and,
3. We consider the generation of the *findings* and *impression* sections of radiology report separately in our proposed model. This helps overcoming challenges arising from concatenating findings and impression section as a single caption, which is one of the limitations of the previous work.

2 Related work

In this section, we present background knowledge about standard deep learning models in medical image classification, image captioning, and summarisation systems, upon which our proposed model is based on. We also present an overview of related work in the radiology report generation from medical images.

2.1 AI in chest radiology

In recent years, due to the remarkable achievements of deep learning [9] in the field of computer vision, more and

more deep learning techniques have been introduced in the field of medical imaging. AI has the potential to transform the radiologists' workflow, carrying a positive knock-on effect for departmental efficiency, as well as improved patient outcomes [10]. Success of deep learning techniques in health care heavily relies on the availability of large-scale annotated medical datasets. Naturally, radiology domain has large-scale medical data in the form of medical images and radiology reports, which is stored in hospital databases such as picture archiving and communication systems (PACS) for reviewing and organising images. The radiological information system (RIS) is also widely used for managing medical imaging data and the associated data. Apart from this, electronic health records (EHRs) are also growing due to wide interest for storing and organising clinical data including notes, pathology, and laboratory data.

Chest radiographs are one of the most widely used medical imaging modality in the world. With the release of large-scale medical datasets such as IU-CXR [2], ChestX-ray14 [11], PadChest [12], MIMIC-CXR [13], and CheXpert [14], research has started toward applying various classification networks in diagnosing pathologies on chest radiographs. Wang et al. [11] applied pre-trained AlexNet [15], GoogLeNet [16], and ResNet-50 [17] to classify eight disease categories. Guendel et al. [18] proposed a local aware dense network for classification of 14 pathology classes in the ChestX-ray14 dataset. Rajpurkar et al. [19] proposed CheXNet, a 121-layer CNN trained on the ChestX-ray14 for the pneumonia detection, which exceeds average radiologist performance. Baltruschat et al. [20] proposed a fine-tuned ResNet-50 network which achieved high accuracy on 4 out of the 14 disease classes in the Chest X-ray dataset. Yao et al. [21] presented a partial solution to constraints in using LSTMs to leverage inter-dependencies among target labels in predicting 14 pathological classes from Chest X-rays.

2.2 Medical image classification

One of the most fundamental tasks in computer vision and pattern recognition is *image classification*. Medical image

classification aims at classifying a biomedical image as normal or abnormal, or assigning multiple disease labels. It may also refer to classifying an abnormality as benign or malignant, or assigning labels showing the severity of a lesion. Two main approaches in medical image classification are: (1) use of conventional hand-crafted features such as local binary pattern (LBP), scale-invariant feature transform (SIFT), speeded up robust features (SURF), and bag of visual words (BoVW); and (2) use of automated features using deep learning techniques. Singh et al. [22] did comparison of conventional features and deep learning features for modality classification and concept detection in medical images. They found that features extracted by CNNs using transfer learning outperform conventional hand-crafted features. In [23], Wang et al. showed that fine-tuning CNN models improves accuracy of CNN model for liver lesion classification. Using pneumonia detection task, Yadav et al. [24] showed that using transfer learning and data augmentation, the accuracy of CNN models can further be improved. Zhang et al. [25] proposed synergic deep learning, which use multiple CNNs simultaneously to mutually learn from each other for medical image classification. Kumar et al. [26] introduced ensemble CNN model for medical image classification. In order to evaluate the performance of existing deep neural network for various medical image classification tasks, Faes et al. [27] did a feasibility study by applying existing deep learning architecture to classify common diseases on five different datasets. Their results demonstrate that these automated algorithms have capability to show performance at par with state-of-the-art methods, and these methods can be used by health care professionals with no coding experience.

2.3 Image captioning

The task of image captioning [8] aims at understanding the image and providing a natural language description of the image as an output. Image captioning is gaining its attention in both computer vision and natural language processing community due to its multi-modal nature. Image captioning is challenging because it not only requires the syntactic and semantic understanding of the language but also need to understand scene type, location, object properties, and their interactions [28]. Image captioning has various applications including interaction of computers with humans, child education, health assistant for elderly care, and providing aid to visually impaired people.

Research in image captioning falls under three main categories, namely *template based*, *retrieval based*, and *generation based* methods. Template-based approaches have fixed templates with a number of blank slots to generate captions. In this approach, first different objects, attributes, and actions are detected and then blank spaces in

templates are filled. Farhadi et al. [29] first extracted triplets of scene elements to fill the template slots for generating image captions. Li et al. [30] proposed the method of first extracting phrases related to detected objects, attributes, and their relationships, and using this information for filling in the templates. To further extend this work, Kulkarni et al. [31] adapted conditional random fields (CRF) to infer objects, attributes, and propositions to fill gaps in sentence templates. Although template-based methods are simple in terms of complexity and can generate grammatically correct captions, these methods do not generate variable-length captions and lack in their capability to generate novel captions. Apart from this, template-based methods have limitations that the rules are usually human-crafted and the generated sentences are usually rigid.

In retrieval-based approaches [29, 32–34], captions are retrieved from a set of existing captions. In this, given an input image, first visually similar images with their captions are found in the training set. The selected captions are called *caption pool* or *candidate captions*. The caption for the query image is selected from the caption pool. Farhadi et al. [29] mapped images and descriptions to a common intermediate space for retrieving captions. Ordonez et al. [34] proposed Im2Text model that utilises global image representations to retrieve and transfer captions from the dataset. Extractive summarisation methods were used to select the best caption for a given image. Hodosh et al. [32] introduced kernel canonical correlation analysis (CCA) to project image and text to a common space for caption retrieval. To alleviate impact of noisy visual estimation, Mason and Charniak [35] first find visual similarity images and retrieve a set of captions. They estimated a word probability density to score the existing captions to select the best caption. Retrieval-based methods produce general and syntactically correct captions. However, they cannot generate image specific and semantically correct captions. Also, retrieval-based methods have limitations in terms of space search complexity due to selecting captions from a large caption pool.

With the rise of deep learning, neural network-based methods become popular due to their tremendous performance on various computer vision and natural language processing tasks. These neural-based methods first analyze the visual content of the image and then generate captions from the visual content using a language model which models the underlying context of captions. Neural-based methods have the ability to generate novel captions that are semantically and syntactically more accurate than template- and retrieval-based methods. Kiros et al. [36] proposed a multimodal language model that captures visual and linguistic context by CNN and LSTM. Inspired by the machine translation research, Vinyals et al. [8] proposed an

end-to-end framework for image captioning by combining CNN and LSTM models. Karpathy and Li [37] proposed a multimodal embedding model that aligns image and text modalities using CNN and LSTM models. Inspired by the natural human behavior to pay attention to specific regions of image and then forming a good relationship of objects in those regions, *attention mechanism* was proposed by Xu et al. [38] that finds salient regions in a image to describe content of an image. Liu et al. [39] proposed novel model with different levels of explicit supervision for learning attention maps during training. This helps improving both the attention correctness and caption quality. It is found that not all words in a caption are associated with visual regions in an image. Hence, semantic concept words that are associated with image regions need to be learned but stop words are not associated with image regions. Various attention mechanism variants that take this into account are proposed by [40, 41]. The traditional image captioning aims at describing an image with only a single sentence. However, one sentence's description is high-level and coarse. By only describing images with a single high-level sentence, there is a fundamental upper-bound on the quantity and quality of information approaches can produce [6, 42]. Thus paragraph generation task is introduced, which aims to describe images in detail with a paragraph rather than a sentence [43].

2.4 Radiology Report Generation

Medical imaging has been an essential tool for clinicians to make diagnosis and deliver treatment. Radiologists routinely examine medical images and describe both normal and abnormal findings in the form of radiology report. Examining medical images is laborious and time-consuming because radiologists have to do lesion measurement and describe morphology for abnormalities. This task of examining medical images and writing textual radiology reports is time consuming, tedious, and often error-prone [44]. A radiology report consists of various sections including *comparison*, *indication*, *findings*, and *impression*. The findings and impression sections are the most important sections of a radiology report. The findings section is a paragraph containing multiple sentences that describe radiologists' observations about normal and abnormal regions present in an image. The impression section is a single-sentence summary or concluding remarks of a radiology study. An automated system that can generate radiology reports from medical images can benefit radiologists to focus more on other high-end complex cognitive tasks. With this ambition, researchers proposed the task of automated radiology report generation from medical images, also known as medical image captioning.

The goal of medical image captioning is to generate accurate, informative, complete, and coherent medical reports. This is a challenging task due to a number of reasons: (1) The generated report should correctly identify all the pathologies or abnormalities present in an image; (2) The generated report should be a long coherent paragraph that provides detailed description about findings in an image; and (3) The generated report should consist of heterogeneous information in the form of sections including *findings* and *impression* [45].

One of the early works integrating visual and linguistic data is toward leveraging radiology reports for medical image annotation in a weakly supervised manner. The detection of subtle disease characteristics using deep learning requires large amount of annotated medical data. Manual annotation of medical data is not scalable due to the requirement of experts, higher cost of annotation, time constraints, and also errors due to the subjective nature of human perception [46]. An alternative approach is to leverage the benefit of large-scale paired medical images-radiology reports data lying in hospitals' PACS. Schlegl et al. [46] proposed the use of semantic content of textual radiology reports linked to the image data instead of voxel-wise annotations for training an image classifier. Their findings indicate that the inclusion of semantic representations got from radiology reports has advantage in improving classification accuracy of pathologies and learning semantic concepts such as spatial position. Shin et al. [47] presented an interleaved text/image deep learning system to extract and mine the semantic interactions of radiology images and reports from hospital's PACS. They mine disease terms from radiology reports using disease ontology and semantics and attach them to medical images and demonstrate the prediction of the presence or absence of disease. Further, Shin et al. [48] used large-scale radiology dataset to mine disease names and to used trained models to infer joint image/text contexts for composite image labeling.

One of the early works toward automated radiology report generation is by [3] where feature-based approach is applied to predict categorical BI-RADS descriptors for breast lesions. Three main descriptors, namely, *shape*, *margin*, and *density*, are used to train a classifier. Further categories for each feature are **shape**: {*oval*, *round*, *irregular*}, **margin**: {*circumscribed*, *indistinct*, *spiculated*, *microlobulated*, *obscured*}, and **density**: {*nonhomogeneous*, *homogeneous*}. Although the trained model can identify these features given a input image and fill them into a fixed template, the approach is limited to keywords and do not support coherent and free-form radiology report. Shin et al. [48] used a cascaded CNN-RNN captioning model to generate description about the detected disease. Their model could generate individual words, however, the

generated words are not coherent and can be difficult to comprehend. This is due to the poor information extraction and language modeling capabilities. Later, Zhang et al. [49] proposed a CNN-RNN model enhanced by an auxiliary attention sharpening (AAS) module to automatically generate medical imaging report. They also demonstrated the corresponding attention areas of image descriptions. Their proposed model can generate more natural sentences but the length of each sentence is limited to 59 words. In addition, the content of a generated report is limited to five topics. Jing et al. [5] proposed a hierarchical co-attention-based model for generating medical imaging report. Their proposed model has capability to attend image features and predict semantic tags while exploring joint effects of visual and semantic information. Also, their model can generate long sentence descriptions in reports by incorporating a hierarchical long-short term memory (LSTM) network. However, they concatenated findings and impression section into a single text for generation task. Also, the experimental results show that their model was prone to generating false positives because of the interference of irrelevant tags.

Li et al. [50] built a hierarchical reinforced agent, which introduced reinforcement learning and template-based language generation method for medical image report generation. Their agent effectively utilises the benefit of retrieval- and generation-based approaches. However, the heavy involvement of pre-processing in extracting templates makes the method heuristic and difficult to generalise to other datasets and applications. Xiong et al. [45] proposed a novel hierarchical neural network-based Reinforced Transformer for medical image captioning (RTMIC) to generate coherent informative medical imaging report. The transformer module speeds up the training process by enabling parallel computing in feed-forward layers. Also, using bottom-up attention using pre-trained DenseNet model on ChestX-ray14 dataset leads to more accurate pathological terminologies in the generated sentences. The use of reinforcement learning-based training method addresses the discrepancy between Maximum Likelihood Estimation (MLE)-based training objective and evaluation metrics of interest. However, only findings section was considered in the study and impression section is not included.

Yin et al. [6] proposed a hierarchical recurrent neural network (HRNN)-based medical report generation model. They also introduced a topic matching mechanism to HRNN, so as to make generated reports more accurate and diverse. The HRNN consists of two RNNs, a sentence RNN and a word RNN. The sentence RNN takes detected abnormalities and the features maps extracted by the CNN as inputs and then generates several topic vectors. Given a topic vector, the word RNN produces an appropriate

sentence. However, the underlying assumption that most sentences are only related a single disease or part of the location of the medical image is wrong. Moreover, they concatenate findings and impression sections as single text and regard the combined text as the ground truth report. Zeng et al. [51] proposed a coarse-to-fine ultrasound image captioning ensemble model that can generate description of ultrasound images. First an organ classifier detects organ present in an ultrasound and then a respective encoder-decoder framework generates report for respective organ.

2.5 Summarisation systems

Automatic text summarisation refers to the task of condensing the documents into a shorter version having the salient information from the source document. Radev et al. [52] define a summary as a text that is no longer than the original text(s) and conveys important information in the original text(s). The reduction of data accomplished by text summarisation allows users to identify and process relevant information more quickly and accurately. In health care, the amount of data in the form of biomedical literature and patient's health records is growing exponentially. It is difficult for clinicians and clinical researchers to cope with this abrupt rise in information. Automatic summarisation systems can definitely help clinicians and researchers to obtain a summary of a long single document or multiple documents [53]. Hence, text summarisation is an important tool to assist clinicians and researchers with their information and knowledge management tasks.

Existing summarisation approaches fall into three categories: *extractive*, *abstractive*, and *hybrid*. Extractive approaches select spans of text from the input text and copy them directly into the summary. Early non-neural techniques utilise domain expertise to develop heuristics for content select [54–56]. On the other hand, more recent neural-based techniques allow for end-to-end training for text summarisation. Abstractive approaches paraphrase the source documents and create summaries with novel phrases not present in the source document [57]. Most of the research in abstractive summarisation use attention and copying mechanism [58–60]. Hybrid approaches are the mix of extractive and abstractive approaches, where first salient sentences are selected using extractive approach and then paraphrasing is done to get novel sentences using abstractive approach [61–64]. To address the limitation that neural models with a fixed vocabulary cannot handle out-of-vocabulary words, a pointer-generator is used to copy elements directly from the input sequence [65]. To overcome the issue of repetition in the generated summaries, a coverage mechanism was proposed [58].

3 Methodology

In this section, first, we define the task and provide a detailed description about our proposed *show, tell and summarise* model. Since annotated medical datasets are of small size, we use pre-trained models that are trained on large-scale annotated datasets to apply knowledge and skills learned from large-scale generic datasets to small size medical datasets.

3.1 Task definition

The first module is an *image classification module*, which classifies an input chest X-ray as normal or abnormal. The image classification module is a binary classification problem where the input is a chest X-ray I and the classifier predicts the labels, $l = \{0, 1\}$, where 0 represents *normal* and 1 represents *abnormal*.

The second module is a *generation module*, which is an *encoder–decoder* framework. It first extracts the global image features using CNN and then fed into an LSTM for generating sequence of words, which forms the *findings* for the input chest X-ray. For generating radiology findings from chest X-rays, we are given an input image I and the model should generate a sequence of words $y = (y_1, y_2, \dots, y_N)$, where y denotes the description in the form of a radiology findings and $y_1 \dots y_N$ denote words in the findings. Given a training set $D = (I, y)$ which consists of (I, y) pairs, where I represents a given medical image and y represents an accompanied radiology findings for the image in the form of $y = (y_1, y_2, \dots, y_N)$, we train the model w.r.t its parameters θ in order to maximise a probabilistic model $p_\theta(y_1, y_2, \dots, y_N | I)$.

The third module is a *summarisation module*, which takes findings generated by the generation module and summarises it to impression as an output. Given a paragraph of findings generated by the generation module which is represented as a sequence of tokens $p = \{p_1, p_2, p_3, \dots, p_N\}$, where N is the number of tokens in the findings section, the summarisation module find a sequence of tokens $q = \{q_1, q_2, q_3, \dots, q_L\}$ that best summarises the salient and clinically significant findings in p , where L is an arbitrary length of the summary or impression section. Since, we do have *findings–impression* pairs in our dataset, we can train a summarisation module in the form of supervised learning.

3.2 Model

The detailed architecture of our *show, attend and summarise* model is shown in Fig. 3.

3.2.1 Image classification module

The goal of the image classification module is to classify an input Chest X-ray into normal or abnormal. The image encoder is a convolutional neural network (CNN) that automatically extracts hierarchical visual features from images. More specifically, we use Inception-v3 [16] model pre-trained on the large-scale ImageNet [66] dataset. We selected Google’s Inception-v3 model for image classification because it has a significantly deeper structure and due to Inception blocks it learns semantically rich representation of images. The repetition of the inception modules captures the optimal sparse multi-scale representations of the images while simultaneously reducing dimensionality. We extract features using Inception-v3 model, which is one of the deep convolutional neural network. We fixed earlier layers of the network to preserve generic features and trained last layers to capture features specific to medical images. The extracted features are passed on to classifier to classify an image as either normal or abnormal. We resize chest X-rays to 229×229 so that we can use pre-trained Inception-v3 model.

3.2.2 Findings generation module

The findings section is a paragraph having high-level description of the image. The goal of findings generation module is to take the global visual features learned by the image encoder as input and generates the findings as an output. For findings generation, a stacked LSTM is used for the decoding step. The LSTM cell has multiplicative gate structure that deals well with exploding and vanishing gradient problem and is a good solution to the problem of long-term sequence dependency. The visual feature vector is used as the initial input of the LSTM to predict the first word of the sentence and then the whole sentence is produced word by word. In our LSTM module, the dimensions of word embedding and the dimensions of hidden states are 300 and 1024, respectively.

In order to generate radiology reports from chest X-rays, we first process data such that given an input image I , the trained model should generate a sequence of words $y = (y_1, y_2, \dots, y_N)$, where y denotes findings of a radiology report and $y_1 \dots y_N$ denote words in the finding section. Given a training set $D = (I, y)$ which consists of (I, y) pairs, we train the model w.r.t its parameters θ in order to maximise a probabilistic model $p_\theta(y_1, y_2, \dots, y_N | I)$.

Our encoder–decoder framework for findings generation is an extension of Vinyals et al. [67]. The framework consists of an encoder which transforms the raw image to a vector representation and a decoder which transforms the encoded image vector into a sequence of words. In a typical encoder–decoder framework, CNN is an encoder for

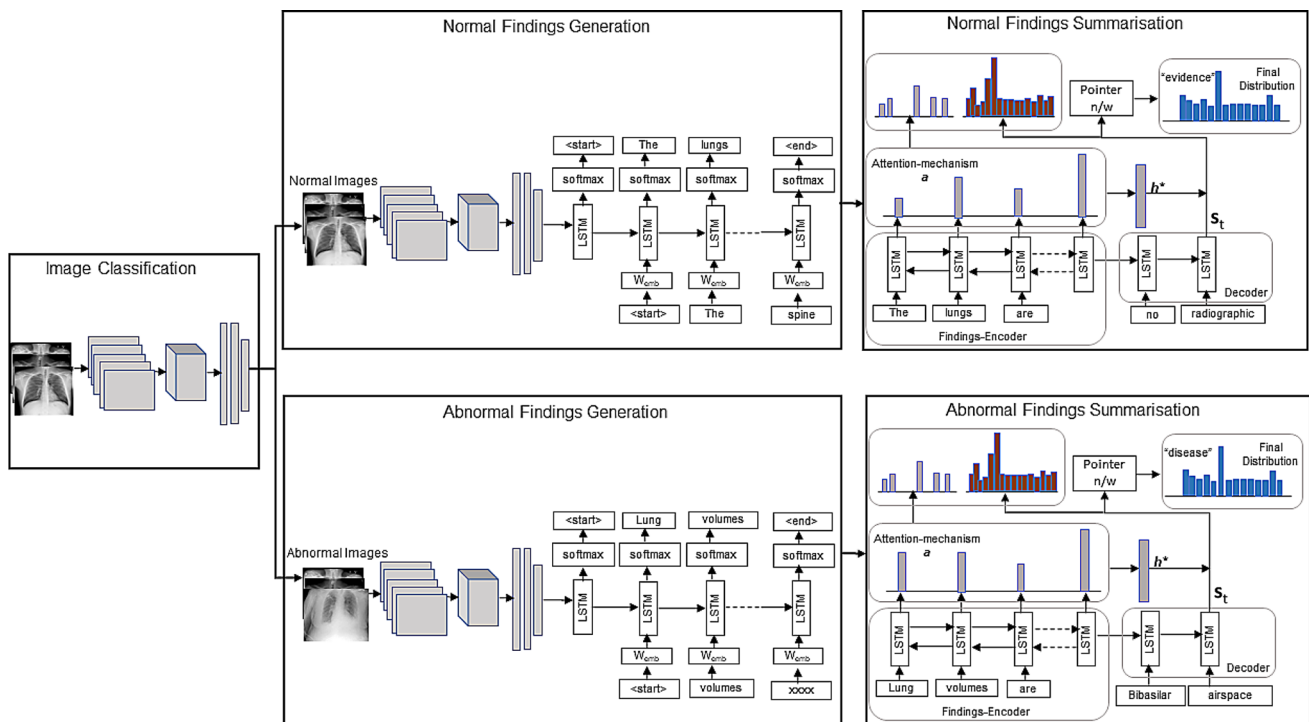


Fig. 3 Architecture of a Show, Tell and Summarise model for radiology report generation from medical images

extracting a global representation of images and RNN is the decoder which models the probability of next words given image and the start token and generates the corresponding sentences. We use Inception-v3 as an encoder and LSTM as a decoder.

3.2.3 Findings summarisation module

The goal of summarisation module is to summarise generated findings by the generation module. We use summarisation module proposed by Zhang et al. [68]. The summarisation module is based on an encoder–decoder architecture that falls under the category of neural summarisation systems. The encoder learns hidden state representations of the input findings and the decoder decodes the input representation into an output sequence in the form of impression.

For the encoder, a bi-directional long short-term memory (Bi-LSTM) network is used. Given the findings sequence in the form $p = \{p_1, p_2, p_3, \dots, p_N\}$, p is encoded into hidden state vectors as given in Eq. 1.

$$h = Bi - LSTM(p), \quad (1)$$

where $h = \{h_1, h_2, \dots, h_N\}$. The last hidden state h_N combines the last hidden states from both directions in the encoder.

After the entire sequence of findings is encoded, the output sequence is generated step by step with a LSTM

decoder. The initial state of decoder is set to the last hidden state, i.e., $s_0 = h_N$. At each t -th step, based on the previous generated token q_{t-1} and the previous decoder state s_{t-1} , the decoder calculates the current state s_t using Eq. 2.

$$s_t = LSTM(s_{t-1}, q_{t-1}), \quad (2)$$

In order to overcome the issue of information loss, an *attention mechanism* [69] which uses a weightage sum of all input states at every decoding step. Given the decoder state s_t and an input hidden state h_i , an input distribution a^t is calculated as:

$$e_i^t = v^T \tanh(W_h h_i + W_s s_t), \quad (3)$$

$$a_i^t = \text{softmax}(e_i^t), \quad (4)$$

where W_h , W_s and v are learnable parameters. A weightage input vectors is calculated as:

$$h_t^* = \sum_i a_i^t h_i, \quad (5)$$

where h_t^* encodes the salient information which is useful in the decoding step t . Finally, an output vocabulary distribution at step t

$$P(q_t | p, q_{<t}) = \text{softmax}(V' \tanh(V[s_t; h_t^*])), \quad (6)$$

where V' and V are learnable parameters.

In order to generate impressions which can leverage “copy” mechanism that can directly copy salient

observations from input findings, a pointer-generator network is used. At each decoding step t , model can decide either to generate a word from the vocabulary with a generation probability p_{gen} , or copy a word directly from the input sequence with probability $1 - p_{gen}$, where p_{gen} is model as:

$$p_{gen} = \sigma(w_{h*}^T * h_t^* + w_s^T s_t + w_q q_{t-1}), \quad (7)$$

where q_{t-1} denotes the previous decoder output, w_{h*} , w_s , and w_q are learnable parameters and σ is a sigmoid function. The overall output distribution in the pointer-generator network is:

$$P(q_t | p, q_{<t}) = p_{gen} P_{vocab}(q_t) + (1 - p_{gen}) \sum_{i: p_i = q_t} a_i^t, \quad (8)$$

where $P_{vocab}(q_t)$ is the same as the output distribution in Eq. (6).

3.3 Transfer learning

Transfer learning [70] defines the ability of a system to recognise and apply knowledge and skills learned in previous tasks to a novel task. In transfer learning, the neural network is trained in two stages: (1) *pre-training*, where the network is generally trained on a large-scale benchmark dataset such as ImageNet [66], representing a wide diversity of labels or categories; and, (2) *fine-tuning*, where the pre-trained network is further trained on the specific target task of interest, which may have fewer labeled examples than the pre-training dataset [71]. The pre-training step helps the network learn general features than can be reused on the target task. Hence, we can use a pre-trained model to detect the basic features using a large-scale annotated dataset and then add our customised layers to detect the data-specific features.

In this paper, we employ *transfer learning* to address the problem of relatively small medical image datasets. We use the pre-trained CNN model trained on the ImageNet dataset to initialise our encoder model. For the language generation and summarisation models, we initialise the LSTM with pre-trained word embeddings trained on large-scale corpora such as *common crawl*, and *Wikipedia*. In order to add medical domain knowledge, we use *Radglove* (Radiology Glove) word embeddings that have been originally trained on 4.5 million radiology reports at Stanford university.

4 Experimental setup

In this section, we first describe about the dataset used for experiments and briefly describe various evaluation metrics to evaluate the proposed model. We then provide details

about several model parameters that are set in model configuration.

4.1 Dataset

In order to evaluate the effectiveness of our proposed methodology, we did experiments on a publicly available dataset comprising medical images and associated reports, namely the Indiana University Chest X-ray collection (IU-CXR) [2]. The dataset is curated from two large hospital systems within the Indian Network for Patient Care databases. The IU-CXR dataset is publicly accessed through the Open Access Biomedical Image Search Engine (OpenI). There are 7470 chest X-rays and 3955 radiology reports. Majority of the studies have both frontal and lateral views of chest X-rays. Each radiology report consists of four sections, namely *Comparison*, *Indication*, *Findings*, and *Impression*. The *comparison* section contains previous information about the patients' study, i.e., any preceding medical examination with which current study is being compared to. The *indication* section contains symptoms of patients or reasons of examination, due to which study is being undertaken. The *findings* section lists the radiology observations. It provides detailed information about normal and abnormal regions in the medical image and its location in the image. The *impression* section summarises the findings section and concludes the radiology report. All reports are fully anonymised using de-identification techniques. About 2.5% of findings and impression words are also removed during anonymisation, resulting in some keywords missing in reports. Since the original data are from multiple hospitals and are inconsistent, there are some images or findings missing in the original dataset.

Our proposed model first generates findings from chest X-rays and then summarises the generated findings to get impression section. Hence, the overall pipeline is as follows: $CXR \rightarrow Findings \rightarrow Impression$. We consider *findings* for the generation task and then passing it to summarisation module to get final *impression* section. In the IU-CXR dataset, majority of reports don't have *comparison* and *indication* sections, so we do not consider them in our study as prior work has also considered the same constraint. For our proposed model, we need both *findings* and *impression* section. We filtered out all reports which do not have both the findings and impression section. With this filtering, we are left with a reduced dataset of 6461 radiology reports. Since the impression section is the most indicative whether a scan is positive or negative and there is no labeled dataset, we applied *rule-based text processing* to classify a radiology report into either *Normal* (Negative) or *Abnormal* (Positive). Out of these, there are 4838 radiology reports that are *Normal* and 1623 are

Abnormal. Table 1 shows the split of our dataset following the same distribution of previous studies [5, 72].

For text processing, we tokenised all text and converted into lower case to map to a common vocab of 1699 words. We removed numbers and special tokens, except full stop. We decided to drop infrequent words with less than 5 frequency. We also added two special tokens, *<start>* and *<end>* to indicate the start and the end of a sentence. To evaluate our models, we randomly shuffle our dataset and dividing into training, development, and testing. The split for the training, development, and testing set is shown in Table 1.

4.2 Evaluation metrics

In order to evaluate the effectiveness of our proposed model, we use standard image captioning metrics, namely BLEU [73], ROUGE [74], METEOR [75], and CIDEr [76], which originated from machine translation and summarisation. In order to check how effective impressions which are summarised from generated findings compared to the reference impressions, we use widely used summarisation ROUGE metric. We report results in the form of F_1 scores for ROUGE-1, ROUGE-2, and ROUGE-L, which measure the word-level unigram overlap, bigram overlap and the longest common sequence between the reference impressions and predicted impressions.

Bilingual evaluation understudy (BLEU), which is a standard evaluation metric for machine translation, and is also widely used in the evaluation of image captioning. BLEU analyzes the correlation of n -grams between the generated text and the reference text. BLEU measures the word n -gram overlap between the generated and the ground truth caption. BLEU-1 considers unigrams, BLEU-2 consider bigrams, and so on. A brevity penalty is added to penalise short generated descriptions. Recall-Oriented Understudy for Gisting Evaluation (ROUGE) counts the number of overlapping units such as n -grams, word sequences, and word pairs between the generated text and the reference text. ROUGE-L is the ratio of the length of the longest common sub-sequences between the machine-

generated text and the reference text. Metric for Evaluation of Translation with Explicit ORdering (METEOR) calculates the harmonic mean of precision and recall. METEOR uses stemming, in particular, Porter stemmer and WordNet ontology to check for synonyms in the text. To take into account longer sub-sequences, it includes a penalty of up to 50% when no common n -grams exist between the machine-generated text and the reference text. METEOR demonstrated to have high levels of correlation with human judgments of translation quality. It has high correlation in the aspects of words, sentences, and subsections with human judgements. Consensus-based Image Description Evaluation (CIDEr) is a novel metric for evaluating image descriptions that uses human consensus. It measures the cosine similarity between n -gram TF-IDF representations of the two captions (words are also stemmed). This is calculated for unigrams to 4-grams and their average is returned as the final evaluation score. All evaluation metrics are implemented in COCO-evaluation code [77].

4.3 Model configurations

All experiments are implemented in Python 2.7 using Keras [78] library with a Tensorflow [79] backend. Since training a deep neural network requires Graphics Processing Units (GPUs) to accelerate the computation process, we use two Nvidia Tesla P100 GPUs on a GPU cluster.

For the *image classification module*, we employed Inception-v3, a CNN model which is already trained on the ImageNet dataset and used it on chest X-rays from the IU-CXR dataset. The Inception-v3 model works as a multi-level feature extractor by computing 1×1 , 3×3 , and 5×5 convolutions. In order to use pre-trained Inception-v3 model, we resized images to 299×299 . Images are normalised between 0 and 1 by dividing the raw pixel values by 255. We use *binary-crossentropy* as the loss metric since we are solving a binary classification problem. We use Adam [80] optimiser with a learning rate of 0.0001.

For the *findings generation module*, we use an encoder-decoder framework which takes classified normal or abnormal chest X-ray as an input and generates normal or abnormal findings, respectively. For the image encoding part, we use Inception-v3 which gives a 2048-dimensional vector as an output. In order to leverage transfer learning by using weights learned on the ImageNet, we process each Chest X-ray having dimensions of 299×299 . For the generation part, the dimension of the LSTM output is set to 300. In order to leverage transfer learning on text data, we use *Glove* embeddings which are 300-dimensional vectors and are given as an input to the LSTM for each word during training. In order to capture the radiology domain context, we use *RadGlove* embeddings providing an embedding of 100-dimension for each word. Instead of

Table 1 Split of normal images and abnormal images into training, validation, and test sets

Data split	Normal # of images	Abnormal # of images
Training	4338	1423
Validation	250	100
Testing	250	100
Total	4838	1623

using a single LSTM as per our baseline model, we incorporate n -stage stacked LSTM, where $n = \{1, 2, 3\}$ to formulate a decoder which helps in learning rich semantic context of radiology reports during training. When initialising decoder with generic Glove embeddings, the dimensionalities of the word embedding space and the hidden state of LSTM are set to 300. For initialising decoder with RadGlove embeddings, the dimensionality of the word embedding space is set to 100 and the hidden state of LSTM is set to 300. The vocabulary is obtained by processing text of findings by turning words to lower case and removing words that are non-alphabet characters and appear less than 5 times. We train our model in an end-to-end manner by minimising our loss function using Adam algorithm with a *batch size* of 32. We set initial learning rate to 0.0005 which decays by a factor of 0.3 after 10 epochs, and the dropout rate is set to 0.2. Early stopping is used to prevent overfitting of the model. The best model configuration is selected based on the highest CIDEr score on the validation set. The experimental results in [67] report beam-search size of $l = 3$ produced the best results. Adapting their approach, during inference, we set beam-search of size 3 to generate radiology report for a given medical image.

For the *summarisation module*, we use an encoder–decoder framework-based neural summarisation which is pretrained on 4.5 million radiology reports at Stanford university [68]. In particular, a neural sequence to sequence model with copy mechanism is used to summarise findings to have conclusive summary or impression of a radiology report.

In order to use transfer learning, Glove [81] algorithm is applied to a corpus of 4.5 million radiology reports. The resulting 100-dimensional word vectors are used to initialise the word embeddings of seq2seq model. A two-layer Bi-LSTM is used for encoders with their hidden size of 100 for each direction. The decoder is a one-layer LSTM with its hidden size of 200. The negative log-likelihood loss is optimised using the Adam optimiser. A dropout of 0.5 is used to avoid the problem of overfitting models. Since impression is generally a single-sentence summary, the decoding is stopped whenever a $\langle EOS \rangle$ token is predicted, otherwise, it get stopped after generating maximum sequence length of 100. We use the official code implementation of summarisation model as available on the github repository of its original authors.¹

Table 2 Experimental results of the image classification module

Precision	Recall	F1-score	Accuracy
0.76	0.61	0.62	0.78

5 Quantitative results

In this section, provide results for three of the modules of our proposed architecture. First, we provide results of binary image classification in terms of Precision, Recall, F1-score, Accuracy, and receiver operating characteristic (ROC) curve. Next, we provide results obtained for generating findings in terms of standard captioning metrics, namely BLEU, CIDEr, METEOR, and ROUGE. We also provide results for the summarisation system comparing the predicted impression and the reference impression, in terms of ROUGE-1, ROUGE-2, and ROUGE-L metrics.

5.1 Results on image classification module

Here, we demonstrate the performance of deep convolutional neural network in classifying normal and abnormal chest X-rays. Table 2 shows results for the image classification module in terms of standard classification metrics, namely Precision, Recall, F1-score, Accuracy, and the area under the ROC (AUC-ROC) score. Figure 4 shows ROC curve for normal/abnormal classifier. During training, the best model is saved based on the accuracy score on the validation set. Based on results in Table 2, the best model gives precision of 0.76, recall of 0.61, F1 score of 0.62, and accuracy score of 0.78 on the hold-out test set.

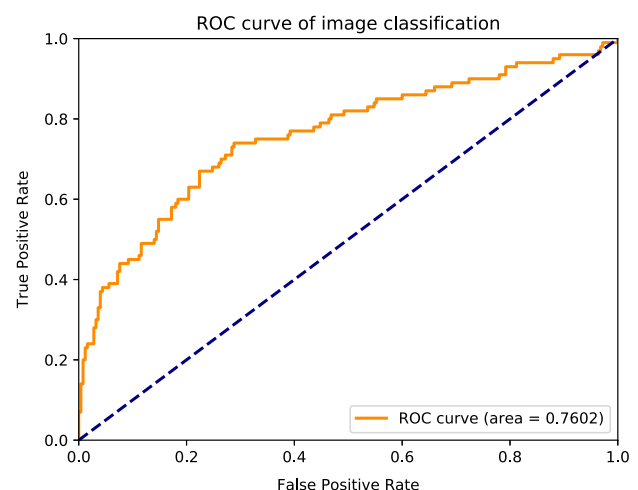


Fig. 4 Receiver operating characteristic curve (ROC) for normal/abnormal image classification

¹ <https://github.com/yuhaozhang/summarize-radiology-findings>.

Table 3 Experimental results on generating radiology findings from normal medical images in terms of BLEU- n ($n = 1, 2, 3, 4$), CIDEr (C), METEOR (M) and ROUGE (R) scores on the test set

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	C	M	R
Show and tell [67]	33.84	21.11	15.15	11.46	60.33	16.46	32.23
Show, attend and tell [38]	35.80	22.54	15.82	11.38	64.37	16.92	32.68
Ours [depth = 1, w/ Glove]	33.32	20.39	14.54	11.00	53.29	16.31	31.54
Ours [depth = 1, w/ RadGlove]	34.69	20.28	13.71	9.71	42.44	16.20	31.28
Ours [depth = 2, w/o Glove]	35.49	21.24	14.64	10.67	49.86	16.52	31.66
Ours [depth = 2, w/ Glove]	35.45	21.50	14.99	11.08	60.08	16.94	32.06
Ours [depth = 2, w/ RadGlove]	35.74	22.67	16.58	12.85	68.23	17.24	32.53
Ours [depth = 3, w/o Glove]	35.90	22.23	16.08	12.48	67.07	16.74	32.44
Ours [depth = 3, w/ Glove]	36.48	22.78	16.50	12.70	73.31	17.21	33.21
Ours [depth = 3, w/ RadGlove]	36.80	22.85	16.82	12.50	75.89	17.68	33.39

The best performance on each metric (column) is shown in bold

w/ denotes model using pre-trained embedding and w/o denotes model without a particular pre-trained embeddings. The proposed model and its variants are compared with baseline model (CNN-RNN)

5.2 Results on generating radiology findings

Table 3 shows results for generating normal radiology findings given a normal chest X-ray to the findings generation module. Prior work considered concatenating findings and impression section as a single text for radiology report generation from chest X-rays. But in our study, we consider generating findings and impression section separately. Hence, no head to head comparison is possible with the previous work in terms of evaluation metrics. We implemented two baseline models, namely *show and Tell* model proposed by Vinyals et al. [8] and *Show, Attend, and Tell* model [38] as baseline models. We compare our proposed model and its variants with the baseline models. The higher value in all metrics represents better generated findings. All the reported results are on the *test* set and are computed by comparing the generated findings with their reference ones as given in our dataset. We observe that our proposed model outperforms the baseline model in all the evaluation metrics with a fair margin. Results show that initialising LSTM decoder with medical pre-trained text embeddings such as *RadGlove* is beneficial compared to random initialisation and initialising LSTM decoder with generic pre-trained embeddings that are trained on large-scale generic text. Since, we aim to generate radiology findings from chest X-rays, *RadGlove* embeddings which are obtained by training Glove algorithm on a corpus of 4.5 million radiology reports is beneficial since it captures the domain knowledge and models the radiological language. It is worth noting that pre-trained Glove embeddings are 300-dimensional word vectors but *RadGlove* pre-trained embeddings are 100-dimensional word vectors. Various studies have found that larger embedding sizes help to capture more context, but still *RadGlove* outperforms Glove embedding highlighting

the importance of incorporating medical domain knowledge for radiology report generation task.

Table 4 shows results for generating reports on abnormal chest X-rays. On comparing results in Tables 3 and 4, we clearly see that scores are lower for abnormal reports in terms of all evaluation metrics. This indicates that abnormal report generation is challenging compared to normal report generation. This is evident from the fact that normal reports are shorter than abnormal reports [72]. Also, if study is normal, radiologists do not provide detailed description in findings section and often summarise their study as normal in the impression section. On the other hand, if study is abnormal, radiologists need to write detailed description about normal and abnormal regions along with their location in chest X-rays. Apart from this, an abnormal radiology report contains many disease terms (or semantic terms) compared to a normal radiology report, making its generation more challenging.

5.3 Results on summarising radiology findings

One of the novelty of our proposed model is that we are generating findings and impression separately, which are sections of heterogeneous information and are required separately as in actual radiology practice. Table 5 shows the results for summarising the generated normal findings to get impression section for normal radiology report. In these set of experiments, we compare the generated summary with the reference summary.² Although, summarisation module takes in generated findings as input text and summarises it to provide impression as the generated text, we provide results for all the variants of the generation module to highlight how effective input findings are to the

² We interchangeably use summary or impression to denote conclusive remarks of a radiology report.

Table 4 Experimental results on generating radiology findings from abnormal medical images in terms of BLEU- n ($n = 1, 2, 3, 4$), CIDEr (C), METEOR (M) and ROUGE (R) scores on the test set

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	C	M	R
Show and tell [67]	27.07	14.92	9.07	5.79	12.02	12.66	24.47
Show, attend and tell [38]	28.98	15.57	9.25	5.95	15.62	12.97	24.72
Ours [depth = 1, w/ Glove]	25.00	13.67	8.55	5.63	10.63	11.72	23.10
Ours [depth = 1, w/ RadGlove]	21.17	12.15	8.04	5.57	15.93	12.24	24.82
Ours [depth = 2, w/o Glove]	26.05	13.49	7.99	5.02	12.85	11.35	21.74
Ours [depth = 2, w/ Glove]	25.45	13.08	7.78	5.10	13.61	11.54	22.12
Ours [depth = 2, w/ RadGlove]	30.62	16.04	9.24	5.63	13.50	13.05	23.23
Ours [depth = 3, w/o Glove]	27.48	13.63	7.34	4.30	9.03	12.08	22.42
Ours [depth = 3, w/ Glove]	24.08	12.28	7.04	4.46	10.44	10.98	21.10
Ours [depth = 3, w/ RadGlove]	23.07	11.86	7.05	4.75	19.78	11.11	23.15

The best performance on each metric (column) is shown in bold

w/ denotes model using pre-trained embedding and w/o denotes model without a particular pre-trained embeddings. The proposed model and its variants are compared with baseline model (CNN-RNN)

Table 5 Experimental results on summarising normal radiology findings in terms of ROUGE-1, ROUGE-2, and ROUGE-L scores on the test set

Method	ROUGE-1	ROUGE-2	ROUGE-L
Show and tell [67]	46.58	24.89	46.33
Ours [depth = 1, w/ Glove]	49.62	27.14	49.28
Ours [depth = 1, w/ RadGlove]	50.04	27.92	49.68
Ours [depth = 2, w/o Glove]	49.73	27.57	49.44
Ours [depth = 2, w/ Glove]	50.79	27.38	50.29
Ours [depth = 2, w/ RadGlove]	47.50	26.34	47.20
Ours [depth = 3, w/o Glove]	48.89	29.86	48.57
Ours [depth = 3, w/ Glove]	49.66	28.33	49.25
Ours [depth = 3, w/ RadGlove]	50.34	29.66	50.07

The best performance on each metric (column) is shown in bold

w/ denotes model using pre-trained embedding and w/o denotes model without a particular pre-trained embeddings. The proposed model and its variants are compared with baseline model (CNN-RNN)

summarisation module, which lead to better summarised impression section. Experiment results show that our proposed model and its variants provide better results compared to the baseline model. From these results, we can also infer that better the generated findings are, the better is the summary from the summarisation module. On comparing results for generating summary of radiology findings on a private dataset at Stanford university [68], we can infer that the summarisation module is effective in summarising radiology findings to provide impression section of a radiology report.

Table 6 shows the results for summarising abnormal findings to get impression section for an abnormal radiology report. From the results, we can infer that our proposed model helps in generating better impression. Comparing the results in Tables 5 and 6, we can see a huge gap in scores of ROUGE-1, ROUGE-2, and ROUGE-L for summarising normal and abnormal radiology findings. The higher scores on summarising normal radiology findings and quite lower scores on summarising abnormal radiology

findings indicate that model is good at generating normal impression section, whereas it shows poor performance in generating abnormal impression section. On digging the reason for this performance, we find that normal impressions are generally a single phrase and there are few fixed templates which are repetitive in majority of normal reports and they are semantically similar in radiology context. For instance, Table 7 shows the most common templates for an impression section in a normal radiology report. These common impression templates for radiology report conclude that the radiology study is normal (Negative). It is important to note that most of these impression templates are single phrase and are semantically similar, which neural seq2seq model can easily learn.

On the other hand, the impression of an abnormal radiology report is rich in diversity in terms of disease terms and is much longer than their normal counterparts. Table 8 shows three representative impression sections of abnormal reports that are selected from the dataset. The first example shows *recommendation of followup study*,

Table 6 Experimental results on summarising abnormal radiology findings in terms of ROUGE-1, ROUGE-2, and ROUGE-L scores on the test set

Method	ROUGE-1	ROUGE-2	ROUGE-L
Show and tell [67]	8.43	1.15	8.03
Ours [depth = 1, w/ Glove]	10.04	2.05	9.65
Ours [depth = 1, w/ RadGlove]	6.35	0.96	5.93
Ours [depth = 2, w/o Glove]	9.75	1.67	9.28
Ours [depth = 2, w/ Glove]	8.81	1.24	8.49
Ours [depth = 2, w/ RadGlove]	10.47	2.23	9.91
Ours [depth = 3, w/o Glove]	6.43	1.11	6.31
Ours [depth = 3, w/ Glove]	7.92	0.60	7.43
Ours [depth = 3, w/ RadGlove]	8.65	1.04	8.20

The best performance on each metric (column) is shown in bold

w/ denotes model using pre-trained embedding and w/o denotes model without a particular pre-trained embeddings. The proposed model and its variants are compared with baseline model (CNN-RNN)

Table 7 Top 15 templates (decreasing order of their frequency) in the impression of normal radiology reports which means the study is *normal*

1	No acute cardiopulmonary abnormality
2	No acute cardiopulmonary findings
3	No acute cardiopulmonary disease
4	No acute cardiopulmonary abnormalities
5	No active disease
6	No acute disease
7	No acute cardiopulmonary process
8	No acute findings
9	No acute radiographic cardiopulmonary process
10	Normal chest
11	No evidence of active disease
12	No acute pulmonary disease
13	No evidence of active disease
14	No acute abnormality
15	Negative for a acute abnormality

which is common if the radiology study is positive and need any additional study to confirm. The second example shows *conversation with physician* to discuss the nature of study. In third impression in Table 8 start with word *stable*, which is temporal in nature, indicating that the current study is compared with any previous study and found that cardiomegaly is stable in nature. Although, such sequential studies are often done in order to see the progression of disease, in the current dataset, this will make language model to learn these words though we don't have any sequential studies, in turn generating spurious radiology reports.

In order to validate the argument of using different models for normal and abnormal cases, we select two of the best model configurations to perform experiments. To train a single model having both normal and abnormal

cases, we combine normal and abnormal images in both training and validation. Since the goal is to check model performance separately on normal and abnormal cases, we keep the same standard split of test set of normal and abnormal cases as per our previous experiments. Table 9 shows results of our ablation studies to see the effect of training separate models. On comparing scores on testing set on both normal and abnormal images having separate models either trained by combining normal and abnormal cases or separating them, we find that scores are higher for separate models in terms of all captioning metrics. This clearly indicates the effectiveness of training separate models for normal and abnormal cases. Table 10 shows results for summarisation radiology findings which are generated by models either by combining normal and abnormal cases otherwise separately. Since $d = 2, w/\text{RadGlove}$ is the best model configuration on the summarisation part for abnormal cases, the scores for model when trained on abnormal cases and tested on abnormal cases are higher than when trained by combining normal and abnormal cases.

5.4 Validation of the show-tell-summarise model on the MIMIC-CXR dataset

In order to validate the generalisation ability of the *show-tell-summarise* model, we did additional experiments on the MIMIC-CXR dataset [13], which contains radiology reports associated with X-ray images.

5.4.1 MIMIC-CXR and MIMIC-CXR-JPG datasets

The MIMIC Chest X-ray (MIMIC-CXR) database v2.0 [13] is a large-scale publicly available chest radiographs in Digital Imaging and Communications in Medicine (DICOM) format with free-text radiology reports. The dataset contains 377,110 chest X-rays corresponding to

Table 8 Sample impression section of representative abnormal reports which mean study is *Positive*

- 1 1. Bullous emphysema and interstitial fibrosis. 2. Probably scarring in the left apex, although difficult to exclude a cavitory lesion. 3. Opacities in the bilateral upper lobes could represent scarring, however, the absence of comparison examination, recommend short interval followup radiograph or CT thorax to document resolution.
- 2 Increased size of density in the left cardiophrenic XXXX. Primary differential considerations include increased size of prominent epicardial fat, pericardial mass, pleural mass or cardiac aneurysm. CT chest with contrast is recommended. These findings and recommendations were discussed XXXX. XXXX by Dr. XXXX XXXX telephone at XXXX p.m. XXXX/XXXX. Dr. XXXX technologist receipt of the results.
- 3 Stable cardiomegaly. Improved aeration of lung bases with persistent left basilar effusion. Prominent interstitium, possibly due to mild volume overload.

XXXX denotes words removed during the de-identification process, resulting in some keywords missing in the report

Table 9 *Generation Ablation Study*: an ablation study to validate the importance of using different models for normal and abnormal radiology report generation

Method	Train	Testing	BLEU-1	BLEU-2	BLEU-3	BLEU-4	C	M	R
Ours [$d = 2$]	Normal + Abnormal	Normal	33.75	20.04	14.00	10.37	45.88	16.00	30.87
Ours [$d = 2$]	Normal	Normal	35.74	22.67	16.58	12.85	68.23	17.24	32.53
Ours [$d = 3$]	Normal + Abnormal	Normal	35.24	21.76	15.54	11.72	64.91	16.60	32.58
Ours [$d = 3$]	Normal	Normal	36.80	22.85	16.82	12.50	75.89	17.68	33.39
Ours [$d = 2$]	Normal + Abnormal	Abnormal	19.95	9.93	5.85	3.62	11.87	10.56	22.28
Ours [$d = 2$]	Abnormal	Abnormal	30.62	16.04	9.24	5.63	13.50	13.05	23.23
Ours [$d = 3$]	Normal + Abnormal	Abnormal	20.25	10.20	6.11	4.13	9.33	10.44	23.14
Ours [$d = 3$]	Abnormal	Abnormal	23.07	11.86	7.05	4.75	19.78	11.11	23.15

Experimental results of generating radiology findings from medical images are in terms of BLEU (B- n), CIDEr (C), METEOR (M) and ROUGE (R) scores on the test set. w/ denotes model using pre-trained embeddings. BLEU- n ($n = 1, 2, 3, 4$) denotes BLEU scores at different n -gram overlap. All these models use *RadGlove* pre-trained embeddings

Table 10 *Summarisation Ablation Study*: an ablation study to validate the importance of using different models for normal and abnormal radiology report summarisation

Method	Training	Testing	ROUGE-1	ROUGE-2	ROUGE-L
Ours [depth = 2]	Normal + Abnormal	Normal	51.18	29.38	50.79
Ours [depth = 2]	Normal	Normal	47.50	26.34	47.20
Ours [depth = 3]	Normal + Abnormal	Normal	52.64	28.76	52.18
Ours [depth = 3]	Normal	Normal	50.34	29.66	50.07
Ours [depth = 2]	Normal + Abnormal	Abnormal	8.96	1.65	8.56
Ours [depth = 2]	Abnormal	Abnormal	10.47	2.23	9.91
Ours [depth = 3]	Normal + Abnormal	Abnormal	8.72	1.50	8.30
Ours [depth = 3]	Abnormal	Abnormal	8.65	1.04	8.20

Experimental results of summarising radiology findings generated from medical images are in terms of ROUGE-1, ROUGE-2, and ROUGE-L scores on the test set. All these models use *RadGlove* pre-trained embeddings

227,835 radiographic studies performed at the Beth Israel Deaconess Medical Center in Boston, MA, USA. The MIMIC-CXR database has chest X-rays with frontal and/or lateral views in DICOM format represented as “.dcm”. Since DICOM format has headers with patient information and stores image data at higher pixel depth and resolution, the database size is of 4.7 TB.

In order to overcome the challenge of handling and comprehending MIMIC-CXR database (DICOM format), the MIMIC-CXR-JPG database [82] is provided. The MIMIC-CXR-JPG database is wholly derived from the original MIMIC-CXR database, and provides 14 structured labels derived from corresponding free-text radiology reports. The source MIMIC-CXR database contains the same images in DICOM format with free-text radiology

reports which were the source of labels. The MIMIC-CXR-JPG dataset provides chest X-rays in JPEG format derived from DICOM files by converting the 16-bit depth raw images into 12-bit depth images in JPEG format.

Since the MIMIC-CXR database provides chest X-rays in DICOM format with their accompanied free-text radiology reports, whereas the MIMIC-CXR-JPG database provides chest X-rays in JPEG format along with structured labels, we took chest X-rays in JPEG format from the MIMIC-CXR-JPG database and free-text radiology reports from the MIMIC-CXR database and did mapping based on study and patient identifiers provided as metadata in both databases. A random subset of 64,000 image-report pairs is selected for this study from the mapped database. We denote this curated dataset as *MIMIC* for the purpose of validating show-tell-summarise model performance and its generalisation ability.

5.4.2 Data processing

The curated *MIMIC* dataset has chest X-rays along with free-text radiology reports which record observations of a practicing radiologist during routine clinical care. The radiology reports are composed of sections including *patient history*, *findings*, and *impressions*. Following the precedent set by the show-tell-summarise model on the IU-CXR dataset, we focus on first generating the *findings* section because it represents the most direct transcription of the imaging study and then summarising the generated *findings* section to produce *impression* section, which concludes the radiology study. There were reports where either *findings* or *impressions* section do not have any textual information. We excluded all reports which do not meet our criteria and thus constrain our dataset to radiology reports having both findings and impression section. The data used in this study are a random subset of the curated *MIMIC* dataset containing 64,000 chest X-rays in JPEG format and radiology reports associated with them having both findings and impressions sections.

To demonstrate *show-tells-summarise* performance and its generalisation ability on the *MIMIC* dataset, we split dataset into training, validation, and testing set based on three experimental settings. The motivation to have three experimental settings is to check model performance with increasing training data, keeping the same test set. This helps us to understand how model behaves with increase in training size. We apply the text classification pipeline to classify radiology report into *normal* or *abnormal* which we applied for the IU-CXR dataset. To ensure balance of both classes and to ensure no data leakage, we curated reduced dataset, having 64,000 image-text pairs keeping 32,000 normal and 32,000 abnormal studies. The distribution of data in terms of training, validation, and testing

Table 11 Statistics for splitting MIMIC-CXR dataset as per three experimental settings

Split	Exp 1	Exp 2	Exp 3
Training	5000	10,000	20,000
Validation	1000	2000	4000
Testing	8000	8000	8000

for each split is shown in Table 11. During the splitting process, we keep same test set for the three experimental settings. This ensures the generalisation ability of the model with increase in samples of training data. Also, if we split data for the largest sample size, i.e., *Exp 3*, we have total 32,000 image-text pairs, providing 20,000 for training, 4000 for validation, and 8000 for testing. We first shuffle the entire dataset and take out test size of 8000 image-text pairs. The remaining data are then split into training and validation based on three experimental settings.

5.4.3 Validation results

Table 12 show results of generating *normal* and *abnormal* radiology findings from chest X-rays for the *MIMIC* dataset. The first column *Setting* denotes three experimental settings for both normal and abnormal radiology findings generation with increasing training size. For each of the three experimental settings, we use same test set while keeping the training and validation size increasing. Except the training size, all hyper-parameters are kept same for three experimental settings. This reduces any bias in model performance due to better hyper-parameter tuning, and demonstrating performance solely based on variation in the data.

The experimental results in Table 12 for both normal and abnormal findings generation show improving model performance with increasing training size. This clearly reflects the nature of deep neural networks as they are data hungry and their performance improvement with increasing training data. On comparing results of show-tell-summarise model for generating normal and abnormal findings given in Tables 3 and 4, respectively, we find model performance on the *MIMIC* dataset at par with the IU-CXR dataset. On comparing results of *normal* and *abnormal* findings generation in Table 12, we can find that the results for abnormal findings generation are low compared to the normal findings generation. This trend is also evident from results for normal and abnormal findings generation on the IU-CXR dataset. Based on the text analysis of radiology reports [72], we hypothesise that there are primarily three reasons for this variation in performance. First, the length of normal radiology reports is shorter than that of abnormal

Table 12 Experimental results of generating radiology findings from medical images on the MIMIC-CXR dataset in terms of BLEU (B- n), CIDEr (C), METEOR (M), and ROUGE (R) scores

Setting	Training size	Type	BLEU-1	BLEU-2	BLEU-3	BLEU-4	C	M	R
Exp 1	5000	Normal	32.22	25.86	21.58	18.48	97.83	15.40	31.94
Exp 2	10,000	Normal	34.02	27.91	23.63	20.44	113.03	16.64	33.84
Exp 3	20,000	Normal	35.38	29.60	25.49	22.34	132.07	17.65	35.32
Exp 1	5000	Abnormal	24.74	15.05	9.57	6.22	39.44	10.13	21.36
Exp 2	10,000	Abnormal	26.62	17.40	11.89	8.36	51.23	11.23	23.64
Exp 3	20,000	Abnormal	27.87	18.93	13.35	9.66	62.19	12.10	25.19

BLEU- n ($n = 1, 2, 3, 4$) denotes BLEU scores at different n -gram overlap. All of these experimental settings use *RadGlove* pre-trained embeddings and have same set of hyper-parameters and test data except the increasing training sample size

radiology reports (IU-CXR data have an average of 32.73 words for normal reports compared to 47.10 for abnormal reports). This is evident from the fact that in actual clinical practice, radiologists often summarise their study in impression section and write quite less if study is normal. On the other hand, for an abnormal study, radiologists provide detailed description of abnormalities indicating which region in the chest X-ray is normal and abnormal. Also, for an abnormal study, radiologists often provide any recommendations of follow-up study. Second, abnormal radiology reports are more diverse and novel compared to normal radiology reports as per our analysis in Table 14. This is evident from the fact that abnormal reports are likely to contain more medical terms, such as disease names, supporting devices, and linking current study to the previous one. On the other hand, normal reports generally have less distinct tokens and often follow fixed templates such as “no acute disease” and “no acute cardiopulmonary findings” (see more examples in Table 7). Third, normal radiology reports often use fixed template phrases which are semantically similar, making neural networks to learn easily. On the other hand, abnormal reports have diversity in phrases, making it hard for neural networks to learn semantics. All of these factors make abnormal report generation more challenging compared to normal report generation.

6 Qualitative results

In this section, we first measure the quality of generated findings and impression sections with their reference counterpart. We also check few random samples in our dataset to check where model is doing good, where model is doing bad, and what is model missing.

6.1 Analysis of generated text

We can manually inspect the quality of generated radiology report in the form of findings and impression section by our

model. However, it is not feasible to do manual evaluation on an entire test set. In order to measure the quality of the generated text, the concept of *diversity* has been explored and it has been found that improving diversity of the generated text leads to more human-like text [62, 83, 84]. Few metrics have been proposed along with standard evaluation metrics that measures the quality of generated captions.

- *vocabulary size ratio*: In this, we calculate the ratio of the number of unique words in the generated text to the number of unique words in the reference text. A higher value indicates that model considers most of words in the reference text during text generation, helping model to have diversity in the generated text. A lower value indicates that method has learnt to do modeling of the most frequent words, ignoring rare words which may be useful in certain applications, such as medical report generation.
- *Novelty*: It is the percentage of generated captions where exact duplicates are not found in the training set. A higher value indicates that model has capability to detect novel objects, its properties, and have ability to generate novel captions.
- *Diversity*: It is the percentage of distinct captions (where duplicates count as a single distinct caption) out of the total number of generated captions within the test set. If out of N generated captions, all captions are distinct, it indicates that model is able to generate diverse captions. On the other hand, if value is lower, it indicates that model has learnt the most frequent templates in the training set, which is one of the problem of current language models and is generating those learnt templates on the test set.

Results of the metrics that measure the quality of the generated captions are shown Table 13. We can see that for a radiology report, the average length of *findings* is quite higher compared to the average length of the *impression* section. The diversity of findings is 18.23% and for

Table 13 Analysis of the complete IU-CXR dataset

Total images	6461
Total sentences in findings	29633
Average sentence length findings	30.88
Vocab size: findings	1687
Unique sentences in findings	5405
Diversity in findings	18.23%
Total sentences in impression	6462
Average sentence length impression	8.64
Vocab size: impression	1261
Unique sentences in impression	1298
Diversity in impression	20.08%

Table 14 Analysis of generated findings and impression by the *normal-d3-radglove* and *abnormal-d2-radglove* models on the test set of the IU-CXR dataset

Criteria	N	A
Unique tokens in reference findings	508	583
Unique tokens in reference impression	221	425
Diversity in reference findings	52.22%	81.76
Diversity in reference impression	34.80%	95.48%
Unique tokens in predicted findings	116	347
Unique tokens in predicted impression	15	166
Diversity in predicted findings	10.48%	60.64%
Diversity in predicted impression	4.00%	50.80%
Novelty in predicted findings	9.49%	17.24%
Novelty in predicted impressions	21.20%	48.00%

We did analysis based on the best model configuration for normal (N) and abnormal (A) findings generation respectively

impression is 20.08%, which indicates that most of the sentences in the entire IU-CXR dataset are repetitive.

We also provide text analysis by comparing the generated findings and impression with their reference counterparts for the best model configurations. Table 14 shows results of comparing generated findings and impression by the *normal-d3-radglove* model, which is the best model configuration for generating normal radiology findings as evident from Table 3. Taking a closer look to these numbers, we can see that there are 508 unique tokens in reference findings for the test set. On the other hand, the findings generated by the model contain only 116 unique tokens. Apart from this, there is a huge gap in terms of diversity in reference text and the generated text. In terms of *novelty*, the predicted impressions have novel sentences that are not seen in the reference text during training, indicating that model is able to generate novel sentences during inferencing.

Table 15 Top 10 sentences in reference findings and predicted findings by the *normal-d3-radglove* model for normal findings generation

#	Reference	Predicted
1	The lungs are clear	The lungs are clear
2	Lungs are clear	Cardiomediastinum silhouette is unremarkable
3	The heart is normal in size	Visualised osseous structures of thorax are without acute abnormality
4	The mediastinum is unremarkable	The lungs are clear bilaterally
5	No pneumothorax or pleural effusion	Specifically no evidence of focal consolidation, pnemothorax or pleural effusion
6	Both lungs are clear and expanded	The cardiomediastinal silhouette is within normal limits for size and contour
7	Heart and mediastinum normal	The heart is normal in size
8	No pneumothorax	The mediastinum is unremarkable
9	There is no pleural effusion or pneumothorax	The lungs are normally inflated without evidence of focal airspace disease, pleural effusion, or pneumothorax
10	Heart size normal	Osseous structures are within normal limits for patient age

Table 14 provides text analysis of generated findings and impression by the *normal-d3-radglove* and *abnormal-d2-radglove* models, which are the best configuration for generating normal and abnormal findings, respectively. In terms of vocabulary size and diversity, the abnormal findings generator provides more diverse findings and impression. On comparing results in Table 14, we can say that the findings generation module and summarisation module for abnormal images use bigger vocabulary size is more diverse and novel.

To see model's capability in generating normal findings, we did analysis of the generated sentences by the model. The top-10 most frequent sentences in reference findings and the predicted findings are shown in Table 15. The sentences in reference findings have smaller length, whereas the predicted sentences are longer, indicating model's ability to compose complex and longer sentences.

The top-10 most frequent sentences in reference abnormal findings and predicted abnormal findings are shown in Table 16. Although the table compares reference text with the generated text for abnormal reports, we can see that most frequent sentences show normal studies. This is because when writing abnormal reports, radiologists provide detailed description about both normal regions and abnormalities present in a chest X-ray.

Table 16 Top 10 sentences in reference findings and predicted findings by the *abnormal-d2-radglove* model for abnormal findings generation

#	Reference	Predicted
1	No pneumothorax	The heart is normal in size
2	The heart is normal in size	No pneumothorax
3	Heart size normal	The mediastinum is stable
4	No pleural effusion or pneumothorax	The mediastinum is unremarkable
5	Heart size is normal	The cardiomedial silhouette is unremarkable
6	Low lung volumes	No acute osseous abnormalities identified
7	No pneumothorax or large pleural effusion	There is no pleural effusion or pneumothorax
8	No focal infiltrates	The heart size is normal
9	The mediastinum is unremarkable	There is no pleural effusion
10	There is no pneumothorax	No pleural effusion or pneumothorax

6.2 Where models are good and bad at?

In order to check where our best configured models are doing good and where they are performing badly, we took representative samples for each model configuration of both normal and abnormal radiology findings and impression modules. Figure 5 shows comparison of baseline model with the proposed model in generating normal findings and impression section for a given chest X-ray, when compared to the ground truth report. On seeing the first example, though both the proposed model and the baseline model provide relevant details in findings, findings generated by the proposed model provide more details such as *focal consolidation* and *osseous structures*. The impression section for both proposed and baseline model is semantically similar, concluding a negative study. In the second example, we can find *scattered calcifications are noted, compatible with prior granulomatous disease* in the ground truth findings, which indicates that radiologist has access to previous study and compared with it to find progression of disease. Both the baseline and the proposed model miss this critical information. The ground truth findings of third example mentions *surgical clips* and *background of marked centrilobular emphysema*, which baseline and the proposed model clearly miss these critical information. Since, there are few example in the training set that shows surgical clips, it is challenging for the model to learn for such a small-scale dataset.

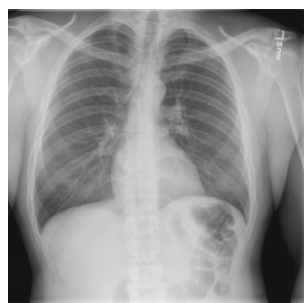
Figure 6 shows qualitative results of the proposed model in generating abnormal findings and impression, when compared to ground truth and the baseline model. In the first case, the ground truth findings show *bilateral interstitial opacities* as an abnormality. The baseline model is missing this critical information but also includes inaccurate or spurious information such as *the aorta is atherosclerotic*, which indicates that the baseline model generates inaccurate or spurious information. On the other

hand, the proposed model is able to highlight *nodular opacity* as given in the ground truth findings. The generated impression section includes *num mm nodular opacity*. This is because the all numbers are mapped to *num* token due to model's incapability to do morphological analysis of lesions. In the second example, again the proposed model is able to correctly identify *low lung volumes*, whereas the baseline model shows normal study. Another notable finding is that generated findings are sometimes redundant, where the model is repetitive in generating most common normal sentences. We also find grammatical errors as well as misplacement of punctuation, which can completely change the meaning of radiology reports. In the last example of Fig. 6, there are two abnormalities: *low lung volumes* and *bibasilar atelectasis versus scarring*. Rest everything is normal in the chest X-ray. The baseline model generates inaccurate abnormality in the form *the mediastinum is stable* and also concludes the study as normal. On the other hand, the proposed model, though is able to correctly identify abnormality *bibasilar airspace opacities*, but also generates spurious information in the form of *the cardiac silhouette is enlarged* and *pleural effusion*.

7 Discussion

The qualitative and quantitative results show that deep learning has potential to automatically generate radiology reports from medical images. With the increasing demand of imaging and improvement in imaging modalities, the role of radiologists is invariably expanding. Artificial intelligence-based clinical decision support systems, such as automatic generation of radiology reports, will allow radiologists to expand their roles including communication with clinicians, leading interventional procedures, conducting quality assurance, and improving health care quality. AI has the potential to develop systems that can

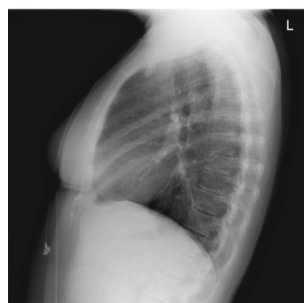
automate routine, tedious, and boring tasks in the health care domain, however, these systems do not have the ability to engage in complex conversation with humans or have empathy similar to humans. Hence, doctors and radiologists will continue to hold an influential and irreplaceable role in patient care, while AI can help enhance such a position. In future, radiologists will help in integrating information from imaging to clinical data as well as offering valuable insights to other medical professionals in the diagnosis and individualised management of patients.



Ground Truth:

Findings: The heart is normal size. The mediastinum is unremarkable. There is no pleural effusion, pneumothorax, or focal airspace disease. The XXXX are unremarkable.

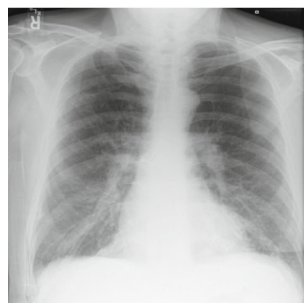
Impression: No acute cardiopulmonary abnormality.



Ground Truth:

Findings: The heart is normal in size and contour. Scattered calcifications are noted, compatible with prior granulomatous disease. The lungs are clear, without evidence of infiltrate. There is no pneumothorax or effusion.

Impression: No acute cardiopulmonary disease.



Ground Truth:

Findings: The XXXX examination consists of frontal and lateral radiographs of the chest. Sternotomy XXXX and surgical clips are again seen. The cardiomeastinal contours are unchanged. There is a background of marked centrilobular emphysema. Streaky opacities in the lung bases may represent atelectasis or scarring. There is no consolidation, pleural effusion or pneumothorax.

Impression: No evidence of acute cardiopulmonary disease or significant interval change.

8 Challenges and future work

In this section, we provide limitations that need to be addressed to have reliable automated systems for radiology report generation. Based on these limitations, we provide future research directions that can help toward generating robust, diverse, and clinically accurate radiology reports.

8.1 Challenges

Although qualitative and quantitative results are encouraging, there are various challenges that need to be

Proposed Model:

Findings: The lungs are clear bilaterally. Specifically no evidence of focal consolidation, pneumothorax or pleural effusion. Cardio mediastinal silhouette is unremarkable. Visualized osseous structures of the thorax are without acute abnormality.
Impression: No evidence of acute cardiopulmonary disease.

Proposed Model:

Findings: Heart size within normal limits. Stable mediastinal and hilar contours. No focal alveolar consolidation. No definite pleural effusion seen. No typical findings of pulmonary edema.
Impression: No acute cardiopulmonary findings.

Proposed Model:

Findings: The heart size and pulmonary vascularity appear within normal limits the lungs are free of focal airspace disease no pleural effusion or pneumothorax is seen.
Impression: No evidence of acute cardiopulmonary disease.

Baseline:

Findings: Lungs are clear. There is no pneumothorax or pleural effusion. The heart and mediastinum are within normal limits. Bony structures are intact.

Impression: No evidence of acute cardiopulmonary disease.

Baseline:

Findings: Heart size and pulmonary vascularity appear within normal limits. The lungs are free of focal airspace disease. No pleural effusion or pneumothorax is seen.

Impression: Normal chest.

Baseline:

Findings: The heart size and pulmonary vascularity appear within normal limits the lungs are free of focal airspace disease no pleural effusion or pneumothorax is seen.

Impression: No evidence of acute cardiopulmonary disease.

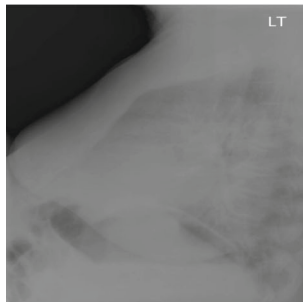
Fig. 5 Qualitative examples of generating *normal findings and impression* sections from chest X-rays. First column denotes chest X-ray given as input to the proposed model. Second column highlights ground truth findings and impression section for the given chest X-ray. Third column denotes generated findings and impression

by the baseline model, and the last column denotes findings and impression generated by the proposed model. XXXX denotes keywords that are removed during data de-identification. Note: The generated text is in lower case. We changed first letter of each sentence to upper case to improve its readability

**Ground Truth:**

Findings: Stable heart size. Diffuse bilateral interstitial opacities. No pneumothorax. No effusions. No acute bony abnormalities.

Impression: Stable bilateral interstitial opacities, could represent scarring or infiltrate. CT may be helpful to further characterize.

**Ground Truth:**

Findings: Lung volumes are low. No focal infiltrates. Pulmonary XXXX are normal.

Impression: XXXX change. Hypoinflation with no visible active cardiopulmonary disease.

**Ground Truth:**

Findings: Low lung volumes. Bibasilar atelectasis versus scarring. Stable left abdominal surgical clips. The heart size and mediastinal silhouette are within normal limits for contour. No pneumothorax or pleural effusions. The XXXX are intact.

Impression: Low lung volumes. Bibasilar atelectasis versus scarring.

Baseline:

Findings: The heart is normal in size. The mediastinum is stable. The aorta is atherosclerotic. There are low lung volumes. No focal consolidation, pneumothorax or large pleural effusion.

Impression: No acute cardiopulmonary abnormality.

Baseline:

Findings: The heart is normal in size. The mediastinum is stable. The aorta is atherosclerotic. There are low lung volumes. No focal consolidation pneumothorax or large pleural effusion.

Impression: No acute cardiopulmonary abnormality.

Baseline:

Findings: The heart is normal in size. The mediastinum is stable. The lungs are clear.

Impression: No evidence of acute cardiopulmonary process.

Proposed Model:

Findings: There is a num mm nodular opacity in the right midlung. No pneumothorax no pleural effusion. No acute bony abnormality.

Impression: num mm nodular opacity in the right midlung.

Proposed Model:

Findings: Low lung volumes bilaterally with central bronchovascular crowding. Without focal consolidation pleural effusion or pneumothoraces. Cardiome-diastinal silhouette is within normal limits. Degenerative changes of the thoracic spine.

Impression: No radiographic evidence of acute cardiopulmonary disease.

Proposed Model:

Findings: The cardiac silhouette is enlarged. There are bibasilar airspace opacities left greater than right with small right pleural effusion. No pneumothorax.

Impression: Bibasilar airspace opacities left greater than right with small right pleural effusion. No pneumothorax .

Fig. 6 Qualitative examples of generating *abnormal findings and impression* sections from chest X-rays. First column denotes chest X-ray given as input to the proposed model. Second column highlights ground truth findings and impression section for the given chest X-ray. Third column denotes generated findings and impression

by the baseline model, and the last column denotes findings and impression generated by the proposed model. XXXX denotes keywords that are removed during data de-identification. Note: The generated text is in lower case. We changed first letter of each sentence to upper case to improve its readability

alleviated before these systems can be deployed in real clinical settings. Various challenges are:

Error propagation Since the proposed methodology has a sequential processing pipeline, an error in the image classification module can be propagated to generation module. Similarly, an error in generation module can be propagated to the summarisation module.

No background information In actual radiology practice, background information in the form of *comparison* and *indication* is available. Due to data anonymisation, most of the background information is lost in the dataset curation process. This limits model's capability to use background information.

No annotated bounding boxes The generic image captioning models improved their performance by using object detection methods based on the annotated bounding boxes

on objects and their relationships. This can be used to ensure that the captioning model pays attention to all the detected objects and their relationships within a scene. On the other hand, medical multimodal datasets lack manually annotated bounding boxes for normal or abnormal regions, making it difficult for the CNN model to get fine-grained features tailored at various medical concepts.

Examining the suitability of NLG metrics for radiology report generation evaluation In order to evaluate the effectiveness of proposed models for radiology report generation, automated evaluation methods are used. Though qualitative human evaluation is possible on fewer samples for error analysis, but manual evaluation on thousands of medical images is not possible. This led to the need of automated methods for radiology report generation using natural language generation evaluation metrics,

mostly inspired from the machine translation and summarisation research. To the best of our knowledge, there is no study undertaken to examine the suitability of automated evaluation metrics for the radiology report generation task. There is no study which reflects to what extent, higher values in terms of evaluation metric highlight better radiology report or do higher values of evaluation metric reflects grammatically correct, diverse, fluent, and clinically accurate report.

8.2 Future work

Based on our text analysis of the generated text, we observe that decoder is generating the most frequent sentences due to limitations of language modeling. To overcome this, in our future work, we aim to use better language models including ELMo [85], transformer [86], and Bidirectional Encoder Representations from Transformers (BERT) [87]. These state-of-the-art language models have capability in interpreting context, polysemous words, nuances, and meanings of the sentences.

9 Conclusions

In this paper, we present a novel *show, tell and summarise* model for radiology report generation from chest X-rays. The proposed model is capable of generating findings and impression section separately, overcoming the limitations of previous studies. The proposed model first classifies an input chest X-ray as normal or abnormal and then respective findings generation module generates the findings. The generated findings are summarised to get an impression section. The proposed model can be a part of clinical decision support systems (CDSS), which augment radiologists by providing “second opinion,” providing draft radiology report, and expedite the clinical workflow. These CDSS systems have potential to fasten clinical workflow, in turn providing early detection of diseases, saving human lives.

Acknowledgements This work is supported by an international Macquarie University Research Excellence Scholarship and the Data61 CSIRO top-up scholarship. This research was undertaken with the assistance of resources from the National Computational Infrastructure (NCI) and the CSIRO Bracwell GPU clusters, supported by the Australian Government.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

- Lewis SJ, Gandomkar Z, Brennan PC (2019) Artificial intelligence in medical imaging practice: looking to the future. *J Med Radiat Sci* 66(4):292–295
- Demner-Fushman D, Kohli MD, Rosenman MB, Shooshan SE, Rodriguez L, Antani S, Thoma GR, McDonald CJ (2016) Preparing a collection of radiology examinations for distribution and retrieval. *J Am Med Inform Assoc* 23(2):304–310
- Kisilev P, Walach E, Barkan E, Ophir B, Alpert S, Hashoul SY (2015) From medical image to automatic medical report generation. *IBM J Res Dev* 59(2/3):2:1–2:7
- Kisilev P, Sason E, Barkan E, Hashoul S (2016) Medical image description using multi-task-loss CNN. In: Carneiro G, Mateus D, Peter L, Bradley A, Tavares JMRS, Belagiannis V, Papa JP, Nascimento JC, Loog M, Lu Z, Cardoso JS, Cornebise J (eds) *Deep learning and data labeling for medical applications*. Springer, Berlin, pp 121–129
- Jing B, Xie P, Xing E (2018) On the automatic generation of medical imaging reports. In: *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long Papers)*. Association for Computational Linguistics, pp 2577–2586
- Yin C, Qian B, Wei J, Li X, Zhang X, Li Y, Zheng Q (2019) Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network. In: *2019 IEEE international conference on data mining (ICDM)*, pp 728–737
- Jing B, Wang Z, Xing E (2019) Show, describe and conclude: on exploiting the structure information of chest x-ray reports. In: *Proceedings of the 57th annual meeting of the association for computational linguistics*. Association for Computational Linguistics, Florence, pp 6570–6580
- Vinyals O, Toshev A, Bengio S, Erhan D (2017) Show and tell: lessons learned from the 2015 mscoco image captioning challenge. *IEEE Trans Pattern Anal Mach Intell* 39(4):652–663
- Goodfellow I, Bengio Y, Courville A (2016) *Deep learning*. The MIT Press, Cambridge
- Lee LIT, Kanthasamy S, Ayyalaraju RS, Ganatra R (2019) The current state of artificial intelligence in medical imaging and nuclear medicine. *BJRiOpen* 1(1):20190037
- Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM (2017) ChestX-Ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: *IEEE conference on computer vision and pattern recognition*. Hawaii, United States, pp 3462–3471
- Bustos A, Pertusa A, Salinas J, de la Iglesia-Vayá M (2019) Padchest: a large chest x-ray image dataset with multi-label annotated reports. [arXiv:1901.07441](https://arxiv.org/abs/1901.07441)
- Johnson AEW, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng CY, Mark RG, Horng S (2019) MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci Data* 6(1):317
- Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, Marklund H, Haghighi B, Ball RL, Shpanskaya KS, Seekins J, Mong DA, Halabi SS, Sandberg JK, Jones R, Larson DB, Langlotz CP, Patel BN, Lungren MP, Ng AY (2019) Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In: *The thirty-third AAAI conference on artificial intelligence, AAAI 2019, the thirty-first innovative applications of artificial intelligence conference, IAAI 2019, the ninth AAAI symposium on educational advances in artificial intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27–February 1, 2019*. AAAI Press, pp 590–597

15. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ (eds) *Advances in neural information processing systems*. Curran Associates, Inc., Red Hook, pp 1097–1105
16. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: *IEEE conference on computer vision and pattern recognition*, pp 2818–2826
17. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: *2016 IEEE conference on computer vision and pattern recognition*, pp 770–778
18. Gündel S, Grbic S, Georgescu B, Liu S, Maier A, Comaniciu D (2019) Learning to recognize abnormalities in chest x-rays with location-aware dense networks. In: Vera-Rodriguez R, Fierrez J, Morales A (eds) *Progress in pattern recognition, image analysis, computer vision, and applications*. Springer, Berlin, pp 757–765
19. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, Ding D, Bagul A, Langlotz C, Shpanskaya K, Lungren MP, Ng AY (2017) Chexnet: radiologist-level pneumonia detection on chest x-rays with deep learning
20. Baltruschat IM, Nickisch H, Grass M, Knopp T, Saalbach A (2019) Comparison of deep learning approaches for multi-label chest X-ray classification. *Sci Rep* 9(1):6381
21. Yao L, Poblentz E, Dagunts D, Covington B, Bernard D, Lyman K (2017) Learning to diagnose from scratch by exploiting dependencies among labels. *CoRR*. [arXiv:1710.10501](https://arxiv.org/abs/1710.10501)
22. Singh S, Ho-Shon K, Karimi S, Hamey L (2018) Modality classification and concept detection in medical images using deep transfer learning. In: *2018 International conference on image and vision computing New Zealand (IVCNZ)*, pp 1–9
23. Wang W, Liang D, Chen Q, Iwamoto Y, Han XH, Zhang Q, Hu H, Lin L, Chen YW (2020) Medical image classification using deep learning. Springer, Berlin, pp 33–51
24. Yadav SS, Jadhav SM (2019) Deep convolutional neural network based medical image classification for disease diagnosis. *J Big Data* 6(1):113
25. Zhang J, Xie Y, Wu Q, Xia Y (2019) Medical image classification using synergic deep learning. *Med Image Anal* 54:10–19
26. Kumar A, Kim J, Lyndon D, Fulham M, Feng D (2017) An ensemble of fine-tuned convolutional neural networks for medical image classification. *IEEE J Biomed Health Inform* 21(1):31–40
27. Faes L, Wagner SK, Fu DJ, Liu X, Korot E, Ledsam JR, Back T, Chopra R, Pontikos N, Kern C, Moraes G, Schmid MK, Sim D, Balaskas K, Bachmann LM, Denniston AK, Keane PA (2019) Automated deep learning design for medical image classification by health-care professionals with no coding experience: a feasibility study. *Lancet Digit Health* 1(5):e232–e242
28. Hossain MZ, Sohail F, Shiratuddin MF, Laga H (2019) A comprehensive survey of deep learning for image captioning. *ACM Comput Surv* 51(6):118:1–118:36
29. Farhadi A, Hejrati M, Sadeghi MA, Young P, Rashtchian C, Hockenmaier J, Forsyth D (2010) Every picture tells a story: generating sentences from images. In: Daniilidis K, Maragos P, Paragios N (eds) *Computer vision—ECCV 2010*. Springer, Berlin, pp 15–29
30. Li S, Kulkarni G, Berg TL, Berg AC, Choi Y (2011) Composing simple image descriptions using web-scale n-grams. In: *Proceedings of the fifteenth conference on computational natural language learning, CoNLL'11*. Association for Computational Linguistics, USA, pp 220–228
31. Kulkarni G, Premraj V, Ordonez V, Dhar S, Li S, Choi Y, Berg AC, Berg TL (2013) Babytalk: understanding and generating simple image descriptions. *IEEE Trans Pattern Anal Mach Intell* 35(12):2891–2903
32. Hodosh M, Young P, Hockenmaier J (2013) Framing image description as a ranking task: data, models and evaluation metrics. *J Artif Int Res* 47(1):853–899
33. Mason R, Charniak E (2014) Nonparametric method for data-driven image captioning. In: *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: short papers)*. Association for Computational Linguistics, Baltimore, pp 592–598
34. Ordonez V, Kulkarni G, Berg TL (2011) Im2text: describing images using 1 million captioned photographs. In: Shawe-Taylor J, Zemel RS, Bartlett PL, Pereira F, Weinberger KQ (eds) *Advances in neural information processing systems*, vol 24. Curran Associates, Inc, Red Hook, pp 1143–1151
35. Mason R, Charniak E (2014) Nonparametric method for data-driven image captioning. In: *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: short papers)*. Association for Computational Linguistics, pp 592–598
36. Kiros R, Salakhutdinov R, Zemel R (2014) Multimodal neural language models. In: Xing EP, Jebara T (eds) *Proceedings of the 31st international conference on machine learning, proceedings of machine learning research*, vol 32. PMLR, Beijing, China, pp 595–603
37. Karpathy A, Fei-Fei L (2017) Deep visual-semantic alignments for generating image descriptions. *IEEE Trans Pattern Anal Mach Intell* 39(4):664–676
38. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhutdinov R, Zemel R, Bengio Y (2015) Show, attend and tell: neural image caption generation with visual attention. In: Bach F, Blei D (eds) *Proceedings of the 32nd international conference on machine learning, proceedings of machine learning research*, vol 37. PMLR, Lille, France, pp 2048–2057
39. Liu C, Mao J, Sha F, Yuille A (2017) Attention correctness in neural image captioning. In: *Proceedings of the thirty-first AAAI conference on artificial intelligence, AAAI'17*. AAAI Press, pp 4176–4182
40. You Q, Jin H, Wang Z, Fang C, Luo J (2016) Image captioning with semantic attention. In: *2016 IEEE conference on computer vision and pattern recognition (CVPR)*, pp 4651–4659
41. Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, Zhang L (2018) Bottom-up and top-down attention for image captioning and visual question answering. In: *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp 6077–6086
42. Krause J, Johnson J, Krishna R, Fei-Fei L (2017) A hierarchical approach for generating descriptive image paragraphs. In: *2017 IEEE conference on computer vision and pattern recognition*, pp 3337–3345
43. Johnson J, Karpathy A, Fei-Fei L (2016) DenseCap: fully Convolutional Localization Networks for Dense Captioning. In: *2016 IEEE conference on computer vision and pattern recognition*, pp 4565–4574
44. Xue Y, Xu T, Rodney Long L, Xue Z, Antani S, Thoma GR, Huang X (2018) Multimodal recurrent model with attention for automated radiology report generation. In: Frangi AF, Schnabel JA, Davatzikos C, Alberola-López C, Fichtinger G (eds) *Medical image computing and computer assisted intervention—MICCAI 2018*. Springer, Berlin, pp 457–466
45. Xiong Y, Du B, Yan P (2019) Reinforced transformer for medical image captioning. In: Suk HI, Liu M, Yan P, Lian C (eds) *Machine learning in medical imaging*. Springer, Berlin, pp 673–680
46. Schlegl T, Waldstein SM, Vogl WD, Schmidt-Erfurth U, Langs G (2015) Predicting semantic descriptions from medical images with convolutional neural networks. In: Ourselin S, Alexander

- DC, Westin CF, Cardoso MJ (eds) Information processing in medical imaging. Springer, Cham, pp 437–448
47. Shin HC, Lu L, Kim L, Seff A, Yao J, Summers RM (2016) Interleaved text/image deep mining on a large-scale radiology database for automated image interpretation. *J Mach Learn Res* 17(107):1–31
 48. Shin H, Roberts K, Lu L, Demner-Fushman D, Yao J, Summers RM (2016) Learning to read chest x-rays: recurrent neural cascade model for automated image annotation. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR), pp 2497–2506. <https://doi.org/10.1109/CVPR.2016.274>
 49. Zhang Z, Xie Y, Xing F, McGough M, Yang L (2017) MDNet: a Semantically and Visually interpretable medical image diagnosis network. In: IEEE conference on computer vision and pattern recognition, Hawaii, United States, pp 3549–3557
 50. Li Y, Liang X, Hu Z, Xing EP (2018) Hybrid retrieval-generation reinforced agent for medical image report generation. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in neural information processing systems, vol 31. Curran Associates, Inc, Red Hook, pp 1530–1540
 51. Zeng XH, Liu BG, Zhou M (2018) Understanding and generating ultrasound image description. *J Comput Sci Technol* 33(5):1086–1100
 52. Radev DR, Hovy E, McKeown K (2002) Introduction to the special issue on summarization. *Comput Linguist* 28(4):399–408
 53. Mishra R, Bian J, Fiszman M, Weir CR, Jonnalagadda S, Mostafa J, Del Fiol G (2014) Text summarization in the biomedical domain. *J Biomed Inform* 52(C):457–467
 54. Neto JL, Freitas AA, Kaestner CAA (2002) Automatic text summarization using a machine learning approach. In: Proceedings of the 16th Brazilian symposium on artificial intelligence: advances in artificial intelligence, SBIA'02. Springer, Berlin, pp 205–215
 55. Filippova K, Altun Y (2013) Overcoming the lack of parallel data in sentence compression. In: Proceedings of the 2013 conference on empirical methods in natural language processing. Association for Computational Linguistics, Seattle, Washington, USA, pp 1481–1491
 56. Colmenares CA, Litvak M, Mantrach A, Silvestri F (2015) HEADS: Headline generation as sequence prediction using an abstract feature-rich space. In: Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: human language technologies. Association for Computational Linguistics, Denver, Colorado, pp 133–142
 57. Kryscinski W, Keskar NS, McCann B, Xiong C, Socher R (2019) Neural text summarization: a critical evaluation. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, pp 540–551
 58. See A, Liu PJ, Manning CD (2017) Get to the point: summarization with pointer-generator networks. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (volume 1: long papers). Association for Computational Linguistics, Vancouver, Canada, pp 1073–1083
 59. Tan J, Wan X, Xiao J (2017) Abstractive document summarization with a graph-based attentional neural model. In: Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers). Association for Computational Linguistics, Vancouver, Canada, pp 1171–1181
 60. Cohan A, Deroncourt F, Kim DS, Bui T, Kim S, Chang W, Goharian N (2018) A discourse-aware attention model for abstractive summarization of long documents. In: Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 2 (short papers). Association for Computational Linguistics, New Orleans, Louisiana, pp 615–621
 61. Hsu WT, Lin CK, Lee MY, Min K, Tang J, Sun M (2018) A unified model for extractive and abstractive summarization using inconsistency loss. In: Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers). Association for Computational Linguistics, Melbourne, Australia, pp 132–141
 62. Liu L, Tang J, Wan X, Guo Z (2019) Generating diverse and descriptive image captions using visual paraphrases. In: 2019 IEEE/CVF international conference on computer vision (ICCV), pp 4239–4248
 63. Gehrmann S, Deng Y, Rush A (2018) Bottom-up abstractive summarization. In: Proceedings of the 2018 conference on empirical methods in natural language processing. Association for Computational Linguistics, Brussels, Belgium, pp 4098–4109
 64. Chen YC, Bansal M (2018) Fast abstractive summarization with reinforce-selected sentence rewriting. In: Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers). Association for Computational Linguistics, Melbourne, Australia, pp 675–686
 65. Moirangthem DS, Lee M (2020) Abstractive summarization of long texts by representing multiple compositionality with temporal hierarchical pointer generator network. *Neural Netw* 124:1–11
 66. Deng J, Dong W, Socher R, Li L, Kai L, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, pp 248–255
 67. Vinyals O, Toshev A, Bengio S, Erhan D (2015) Show and tell: a neural Image Caption generator. In: 2015 IEEE conference on computer vision and pattern recognition, pp 3156–3164
 68. Zhang Y, Ding DY, Qian T, Manning CD, Langlotz CP (2018) Learning to summarize radiology findings. In: Proceedings of the ninth international workshop on health text mining and information analysis. Association for Computational Linguistics, Brussels, Belgium, pp 204–213
 69. Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: 3rd International conference on learning representations, ICLR 2015; Conference date: 07-05-2015 Through 09-05-2015
 70. Razavian AS, Azizpour H, Sullivan J, Carlsson S (2014) CNN features off-the-shelf: an astounding baseline for recognition. In: 2014 IEEE conference on computer vision and pattern recognition workshops, pp 512–519
 71. Raghu M, Zhang C, Kleinberg J, Bengio S (2019) Transfusion: understanding transfer learning for medical imaging. In: Wallach H, Larochelle H, Beygelzimer A, AlcheBuc F, Fox E, Garnett R (eds) Advances in neural information processing systems, vol 32. Curran Associates, Inc, Red Hook, pp 3347–3357
 72. Singh S, Karimi S, Ho-Shon K, Hamey L (2019) From chest x-rays to radiology reports: a multimodal machine learning approach. In: 2019 digital image computing: techniques and applications (DICTA), pp 1–8
 73. Papineni K, Roukos S, Ward T, Zhu WJ (2002) BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the association for computational linguistics. Philadelphia, Pennsylvania, United States
 74. Lin CY (2004) ROUGE: a package for automatic evaluation of summaries. In: 42nd Annual meeting of the association for computational linguistics. Barcelona, Spain, pp 1–8
 75. Banerjee S, Lavie A (2005) METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. Ann Arbor, Michigan, United States, pp 65–72

76. Vedantam R, Zitnick CL, Parikh D (2015) CIDEr: consensus-based image description evaluation. In: IEEE conference on computer vision and pattern recognition. Boston, Massachusetts, United States, pp 4566–4575
77. Chen X, Hao Fang TYL, Vedantam R, Gupta S, Dollár P, Zitnick CL (2015) Microsoft COCO captions: data collection and evaluation server. [arXiv:1504.00325](https://arxiv.org/abs/1504.00325)
78. Chollet F et al (2015) Keras. <https://github.com/fchollet/keras>
79. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M et al (2016) Tensorflow: a system for large-scale machine learning. In: Proceedings of the 12th USENIX conference on operating systems design and implementation, OSDI'16. USENIX Association, USA, pp 265–283
80. Kingma DP, Ba J (2015) Adam: a method for stochastic optimization. In: Bengio Y, LeCun Y (eds) 3rd International conference on learning representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings
81. Pennington J, Socher R, Manning CD (2014) GloVe: global Vectors for word representation. In: Empirical methods in natural language processing. Doha, Qatar, pp 1532–1543
82. Johnson AEW, Pollard TJ, Greenbaum NR, Lungren MP, Deng CY, Peng Y, Lu Z, Mark RG, Berkowitz SJ, Horng S (2019) MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs
83. Lindh A, Ross RJ, Mahalunkar A, Salton G, Kelleher JD (2018) Generating diverse and meaningful captions. In: Kůrková V, Manolopoulos Y, Hammer B, Iliadis L, Maglogiannis I (eds) Artificial neural networks and machine learning—ICANN 2018. Springer, Cham, pp 176–187
84. Deshpande A, Aneja J, Wang L, Schwing AG, Forsyth D (2019) Fast, diverse and accurate image captioning guided by part-of-speech. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 10687–10696
85. Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L (2018) Deep contextualized word representations. In: Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long papers). Association for Computational Linguistics, New Orleans, Louisiana, pp 2227–2237
86. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Lu, Polosukhin I (2017) Attention is all you need. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in neural information processing systems, vol 30. Curran Associates, Inc, Red Hook, pp 5998–6008
87. Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). Association for Computational Linguistics, Minneapolis, Minnesota, pp 4171–4186

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.