Pattern Recognition 114 (2021) 107856

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Automatic medical image interpretation: State of the art and future directions

Hareem Ayesha[a], Sajid Iqbal[a,*], Mehreen Tariq[a], Muhammad Abrar[b], Muhammad Sanaullah[a], Ishaq Abbas[a], Amjad Rehman[c], Muhammad Farooq Khan Niazi[d], Shafiq Hussain[e]

[a] Department of Computer Science, Bahauddin Zakariya University, Multan, Pakistan
[b] Department of Computer Science, Muhammad Nawaz Shareef University of Agriculture, Multan, Pakistan
[c] AIDA LAB CCIS Prince Sultan University Riyadh Saudi Arabia
[d] Bakhtawar Amin Memorial Trust Hospital, Multan, Pakistan
[e] University of Sahiwal, Sahiwal, Pakistan

ARTICLE INFO

Article history:
Received 10 March 2020
Revised 10 November 2020
Accepted 26 January 2021
Available online 29 January 2021

Keywords:
Attention mechanism
Automatic captioning
Convolutional neural network (cnn)
Deep learning
Encoder-decoder framework
Image captioning
Long-Short-Term-Memory (LSTM)
Medical image caption

ABSTRACT

Automatic Natural language interpretation of medical images is an emerging field of Artificial Intelligence (AI). The task combines two fields of AI; computer vision and natural language processing. This is a challenging task that goes beyond object detection, segmentation, and classification because it also requires the understanding of the relationship between different objects of an image and the actions performed by these objects as visual representations. Image interpretation is helpful in many tasks like helping visually impaired persons, information retrieval, early childhood learning, producing human like natural interaction between robots, and many more applications. Recently this work fascinated researchers to use the same approach by using more complex biomedical images. It has been applied from generating single sentence captions to multi sentence paragraph descriptions. Medical image captioning can assist and speed up the diagnosis process of medical professionals and generated report can be used for many further tasks. This is a comprehensive review of recent years' research of medical image captioning published in different international conferences and journals. Their common parameters are extracted to compare their methods, performance, strengths, limitations, and our recommendations are discussed. Further publicly available datasets and evaluation measures used for deep-learning based captioning of medical images are also discussed.

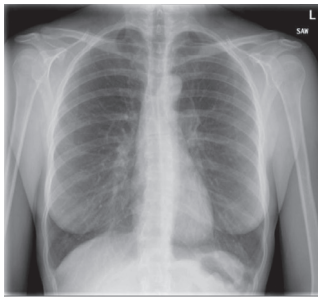
1. Introduction

Automatic image caption generation is a task of extracting the contents of an image through different feature extraction techniques and describing those contents through natural language sentences using natural language processing (NLP). Image captioning is the combination of two artificial intelligence fields that include computer vision to extract visual representations and natural language processing to explain that representations in simple English like sentences. This is a challenging task that goes beyond object detection, segmentation, and classification because it also requires the understanding of the relationship between different objects of an image and the actions performed by these objects as visual representations and to convert these representations into English like sentences. With the availability of large datasets, most widely used approaches for image captioning based on machine learning methods are gaining popularity day by day. Image captioning is helpful in many tasks like helping visually impaired persons, information retrieval, early childhood learning, producing human like natural interaction between robots, and many more but in the medical imaging field this topic has yet to gain popularity because this field has its own problems.

In the medical sector, use of medical images is ubiquitous, for example, medical professionals and radiologists use medical images for diagnosing and treatment of diseases. Pharmacists may use them for drug discovery and surgeons may use imaging in pre-operational, post-operational, and during the operation to monitor the treatment process. Competent medical professionals

* Corresponding author. Assistant Professor, Department of Computer Science, Bahauddin Zakariya University, Multan, Pakistan.
E-mail addresses: sajid.iqbal@bzu.edu.pk, sajidiqbal@bzu.edu.pk (S. Iqbal).

https://doi.org/10.1016/j.patcog.2021.107856
0031-3203/



Impression: Emphysema, however no acute cardiopulmonary finding.

Findings: The heart size and cardiomediastinal silhouette are normal. There is hyper-expansion of the lungs with flattening of the hemi-diaphragms. There is no focal airspace opacity, pleural effusion, or pneumothorax. There multilevel degenerative changes of thoracic spine.

MTI Tags: degenerative change

Fig. 1. Illustration of Chest X-Ray report.

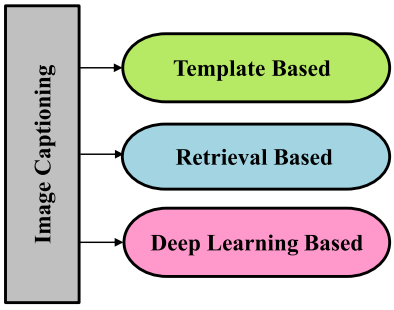


Fig. 2. Image captioning approaches.

manually write textual reports after examining these medical images containing their findings (normal, abnormal, and potentially abnormal) in full paragraphed descriptions as shown in Fig. 1.

For inexperienced or less experienced examiners, writing medical report in textual form may be error-prone because it requires deep understanding of the disease, medical imaging, and thorough analysis of images under consideration. Also, for experienced medical professionals, this task is time-consuming and laborious because it takes at least half an hour to examine an image and write their findings in the form of a report whereas they have to examine many medical images per day. So, it is an unpleasant and tiresome work to do for both experienced and inexperienced medical professionals. Due to shortage of medical professionals in a country like Pakistan that has a huge population, work overload increases tremendously and in areas with the shortage of medical treatment facilities, the proportion of wrong diagnosis is higher [1]. Pakistan is a low-income country where it is an extra expense that comes in the form of an extra fee of their visit to doctors again for asking about their problems/prescriptions in the form of a report.

To facilitate the medical image reporting process, many computer-aided report generation systems based on an image captioning are proposed that automatically extract the findings from medical images and generate a textual report containing fine-grained information like an expert doctor. This saves time of doctors, consumed in manually extracting features from images and then writing a textual report so their workload can be reduced. Moreover, it also helps to reduce the requirement of extra professionals to write reports; the whole process of a medical report generation is automatic and efficient.

The generated reports can be used by many potential users. For example, radiologists can use this report for cross checking, monitoring the subtle changes, and in final decision making. It can also be used for the second opinion by peer doctors. It can be used by technologists to find immediate and first-hand information about an image under consideration. It can also be used in case of emergency in case where expert doctors are not available at the moment. So, this automatically generated report can be used as a context for further treatment without waiting for an expert doctor to generate the report.

But, this also has a variety of challenges like in generating medical reports, instead of a single sentence caption; we have to generate large paragraph that is a non-trivial task. Moreover, a medical report contains heterogeneous information (Fig. 1), for example, a text description (radiologists narrate their observations in this section), an impression (a diagnosis is provided in this sentence), comparison, and a list of tags (keyword from findings containing of critical information). So, to use all this heterogeneous information in generating sentences multiple stages are involved that may include pre-processing, segmentation, feature selection, feature extraction, and classification. Segmentation of different regions of interest is another challenging task because some modalities of medical imaging like ultra-sound contains a lot of noise [2], so identification of regions possessing abnormality is difficult. Another challenge is the limited availability of quality datasets in the medical field. Researchers have developed datasets containing of medical images to excel research in this field that may include IU Chest X-Ray [3], Chest X-Ray14 [4], PEIR Gross [5], BCIDR [6], CheXpert [7], MIMIC–CXR [8], PadChest [9] and ICLEFCaption [10,11].

Image captions can be generated using several approaches that can be broadly classified into three categories as in Fig. 2.

In template-based method, first objects and attributes are detected then captions are generated following specified grammar rules and constraints or through sentence templates. Generated captions are very small and grammatically correct, but, the disadvantage of this approach is that generated captions are hard-coded having no variety and flexibility in them. The second approach is retrieval-based in which new images similar to input image are retrieved from dataset along with their captions also. The new generated caption of input image is either the same caption of most similar image retrieved or the combination of many candidate captions. The third approach relies on deep learning (DL) based neural networks (NN) to generate automatic captions. In this method network is trained on end to end mapping from images to captions. Template-based and retrieval-based approaches are early work, but now-a-days, state-of-the-art is deep neural networks that are widely used in the medical image description generation. The network architecture used for this purpose may include an encoder-decoder framework, fully connected networks, and convolutional networks. Encoder is a convolutional neural network (CNN) that extracts the visual features from images in hierarchical manner and can be trained directly on the application dataset in hand or used as a pre-trained model such as VGGNet [12], ResNet [13], and Inception-V3 [14]. Decoder is a language generating module that is a recurrent neural network (RNN) [15] or it's variant such as gated recurrent unit (GRU) [16] and long short-term memory (LSTM) [17], and generates natural language captions. Recently, an attention mechanism is introduced that lies between encoder and decoder and is used to give importance to salient parts of images corresponding to which captions are generated.

No classic machine learning has been employed in medical image captioning because in ML only a limited number of features are extracted manually that is a difficult task and not good enough to produce good results. On the other hand, medical images are very complex and DL based techniques can handle such challenges and complexities occurring during the generation of medical image captions. So, in the last three years, medical image captioning based on DL has gained a lot of attention and many papers are published in this area. Still there are some problems of using DL for image captioning. For example, DL-based models (i.e. CNNs) require a large amount of training data to avoid overfitting problem and to improve the generalizability of the model. However, due to the scarcity of such large scale publically available datasets,

it is challenging to train new deep models from scratch. Transfer learning-based approaches come here as rescue to perform this task. Secondly, language models (LSTM) consumes high computation power and training time because of their sequential nature and also suffer from vanishing gradient issue.

According to best of our knowledge, only one survey paper [18] has been published on this topic. Although, a good literature survey [18] is presented in his paper, but it is not structured and coherent. In our work, we are aimed to provide a comprehensive and structured review of automatic captioning for medical images generated using different imaging modalities. Our major focus is deep-learning based approaches and minor focus on retrieval-based methods that are using deep neural networks to generate medical image captions. The rest of the paper is organized in the following manner: In Section 2, some tasks in medical image analysis using deep learning are described. Section 3 provides a summary of our study methodology. In Section 4, a brief introduction of publicly available datasets used for medical image captioning is given. Section 5 provides details about evaluation measures used for deep-learning based image captioning. In Section 6, medical image captioning methods are categorized in different ways. In Section 7, reviewed methods are compared on different datasets used by researchers. Our findings of reviewed studies and some potential future directions are discussed in Section 8. Finally, Conclusion and our future work are described in Section 9.

2. Deep learning in medical imaging

A number of Deep Learning (DL) methods are being used to perform various medical image analysis tasks [19]. Researchers are experimenting methods, designed to perform different tasks, for a medical image description generation too. In addition, these methods are also being used for medical video captioning, enhancing the resolution of 2D and 3D medical images, medical image generation, data completion, discovering patterns, removing obstructing objects in a medical image, and normalizing a medical image. Some other widely used medical imaging DL based tasks related to our work are medical image classification, retrieval, object detection, medical concepts prediction from medical images, and finally medical image caption generation. These methods are reviewed in the following text.

2.1. Image classification

Medical image classification is one of the first tasks in which deep learning models were used. In this task, medical images are classified as either normal or abnormal and further classification of abnormal images, for example, classification among different tumor types [20,21]. It also includes assigning different disease labels to images [22,23] for example, grading tumor based on the severity of disease in image. Medical image classification can be done using pre-trained networks, fine-tuning the pre-trained networks or training one's own network from scratch. Most widely used networks for medical image classification are CNN's for training from scratch or transfer learning. Image classification can also be used in the process of generating captions of medical images. For example, generating textual descriptions against classified abnormal images [24,25] or all disease labels assigned to images [2].

2.2. Object detection

Another major task using CNN's and other network types is object detection in medical images. It includes detecting the location of abnormal organs, regions, and highlighting the abnormalities using a bounding box and localizing landmarks. Exact shape of the organs, objects or regions is not known in this task as the shape of bounding box can be either square or rectangular. Li et al. [26] used Faster RCNN for detecting blurry and smaller regions having cancer. Object detection can be the first step to assist other tasks using medical images, for example, classification or segmentation [22,27], and captioning against detected objects rather than the whole image [2]. Many computer aided detection (CAD) systems are built to localize, detect the abnormal regions increasing diagnosis accuracy, and reducing the doctor's time [28,29].

2.3. Image retrieval

Medical image retrieval is the process of discovering the medical images from massive medical databases using certain features like disease, symptoms, and other cases of medical. Cai et al. [30] used CNN and Qayyum et al. [31] used deep CNN for content based medical image retrieval (CBMIR) systems. Extracting these effective features from the pixels of an image and associating them with some labels or concepts is a challenging task. Deep CNN networks can handle this problem by extracting the high level and rich features. CBMIR is also being incorporated with computer diagnosis systems to assist radiologists in making diagnosis decisions [32]. Medical image captions can also be generated using this task of retrieving the most similar image to input image, and associating the caption of retrieved image to input image captions [33,34].

2.4. Medical concepts prediction

Unified Medical Language System (UMLS) is the collection of files containing biomedical and health vocabularies. UMLS metathesaurus is one of the base knowledge sources of UMLS. It contains concepts, meanings, and concepts names. All this data is collected from different resources in the form of Concept Unique Identifiers (CUIs) [35]. A CUI comprises of 8 characters, starting with "C", and is followed by 7 digits. In medical imaging, medical concepts prediction is the task of predicting CUIs of UMLS metathesaurus giving medical images [10,11]. In this task, first visual features are extracted from images and then concepts are predicted from visual feature vector. These concepts can be used for generating further full sentence captions [33].

2.5. Medical image caption generation

It is the task of generating natural language description against a medical input image. In DL-based captioning, an input image is given to an encoder usually CNN that encodes the image into fixed length vector representation. This vector is fed into a decoder usually LSTM based network part [17] that generates the natural language description. Both Single and multiple-sentence captions of different length can be generated in this task. A key factor that differentiates it from medical concepts prediction is that, medical image captioning includes another AI field named as natural language processing (NLP). In this work, we are presenting a detailed review on DL-based medical image captioning.

3. Study methodology

In order to collect the state of the work on our topic, we have explored different research search engines, conferences, and high-quality journals. The search engines include IEEE Xplore, refseek, Virtual LRC, ACM digital library, scinapse, Google Scholar, Elsevier Science Direct, and Springer Link search engines. Other source examples like conferences and journals include Pattern Recognition, IEEE conference on computer vision and pattern recognition (CVPR), IEEE Access, Journal of the American Medical Informatics Association, neural information processing systems (NeurIPS),

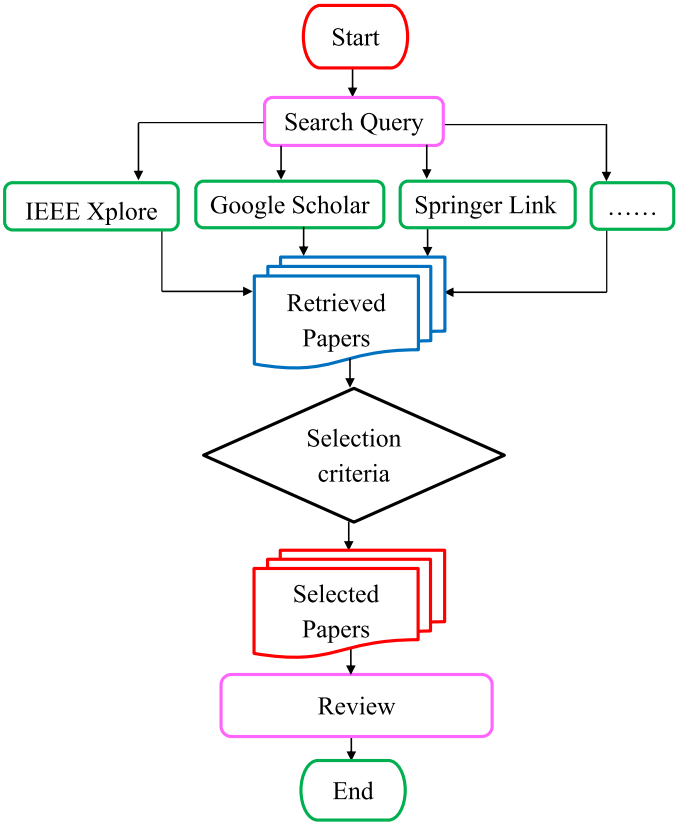
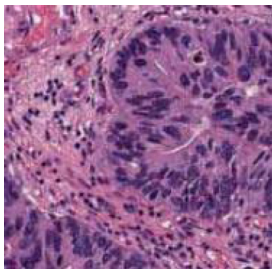


Fig. 3. flowchart for review process.

International Conference on Security, Pattern Analysis, and Cybernetics (SPAC), International conference on machine learning, Conference on Medical Image Computing and Computer-Assisted Intervention, and International Conference on Image Processing (ICIP). Keywords used for search from above mentioned databases include "image caption", "Image description generation", "image captioning", "automatic image captioning", "medical image captioning", "medical report generation", and "captioning using medical images". This built a large list of publications which require the filtering process to sort out low quality works. The designed research papers' selection criteria are 1) papers and studies from 2017 to 2020 are included 2) papers from renowned researchers 3) papers produced in high rank journals and conferences 4) preference is given to the papers that made their source code available publicly. Such papers are useful to recreate the same results for verification of the proposed methodology. 6) Research that is based on publicly available datasets 7) and research work that has got large number of citations. The procedure adopted to conduct this study is shown in Fig. 3. Selected papers are further categorized in Section 6.3.

4. Datasets

Datasets for medical image captioning consists of medical images and corresponding descriptions. These descriptions may be comprised of a single sentence or multiple sentences in the form of a medical report. Only a limited number of datasets for medical image captioning are publicly available. This include IU Chest X-Ray [3], Chest X-Ray14 [4], PEIR Gross [5], BCIDR [6], CheXpert [7], MIMIC–CXR [8], PadChest [9] and ICLEFCaption [10,11] that are described in detail in the following text.



Severe pleomorphism is present in the nuclei. **The nuclei are crowded to a moderate degree**. Basement membrane polarity is partially lost. **Mitosis is infrequent through the tissues**. The nucleoli are mostly inconspicuous**. High**

Fig. 4. Illustration of histopathology report.

4.1. ICLEFCaption

Image Cross Language Evaluation Forum (ImageCLEF) is an evaluation campaign that offers different research tasks which change from year to year (https://www.imageclef.org/). ImageCLEF released ICLEFCaption dataset in 2017 for the ImageCLEFcaption task which consists of two sub-tasks: concept prediction and caption generation. The dataset is publicly available and can be accessed after registration. It consists of 184,614 images extracted from PubMed Central (PMC) (https://www.ncbi.nlm.nih.gov/pmc/). The original caption and set of Concept Unique Identifiers (CUI) extracted through Quick-Unified Medical Language System (UMLS) [36] are associated with each image. This dataset contains multiple-sentence captions and a large number of clinical images also including compound images.

There are multiple shortcomings in this dataset that include inconsistency in dataset; images are highly diversified, there are duplicate captions, captions are of different lengths and many captions have no associated tags with them [10]. This dataset contains 10–20% overall noise because of compound images and the extraction of tags through probabilistic method; UMLS. In 2018, planners of the ImageCLEFcaption extracted clinical /radiology images from dataset of 2017 and also removed compound images to remove noise. This made a homogeneous dataset of 232,305 images-caption pairs. For the concept prediction task, 111,555 concepts were extracted through quick-UMLS [11] from 222,305 images. Nearly 30 UMLS CUI's are associated with each image. Although, many compound images are removed, the resultant dataset is still noisy [18].

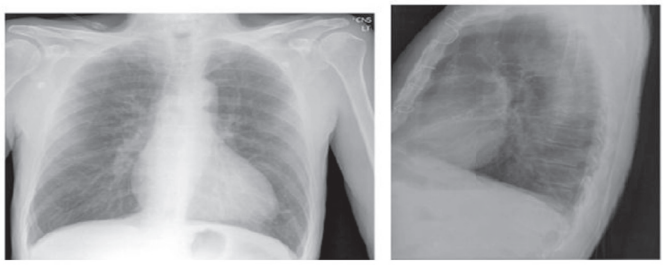
4.2. BCIDR

The Bladder Cancer Image and Diagnostic Report (BCIDR) dataset is comprised of 5000 images with the corresponding report [6]. It was collected with the help of pathologists. Pathologist described observation considering five types of features where each feature describes the appearance of a cell. Observations are followed by a conclusion statement composed of four classes. The same procedure was done by four non-experts in bladder cancer who described observations in their own language, but considering pathologist's reports to assure accuracy. In this way, total five ground-truth reports exist against each image. Each report's length is between 30 and 59 words. A sample image with its diagnostic report is shown in Fig. 4 below:

4.3. IU chest X-Ray and chest X-Ray14

Demner-Fushman et al. [3] presented a collection of chest x-ray images and corresponding radiology reports consisting of findings in textual form. The datasets contains 7470 chest x-ray images including both frontal and lateral views (Fig. 5).

Against each image, there is a textual radiology report that contains five sections. Impression narrates the final diagnosis or



Comparison: Chest radiograph from XXXX.XXXX
Indication: Chest pain
Findings: The cardiac silhouette is border line enlarged. Otherwise there is no focal opacity. Mediastinal contours are within normal limits. There is no large pleural effusion. No pneumothorax.
Impression: Borderline enlargement of the cardiac silhouette without acute pulmonary disease.

Fig. 5. Frontal and lateral chest X-Rays with an associated report (IU Chest X-Ray).

Tags:
MTI: Calcified granuloma
Manual: Calcified Granuloma/lung/bilateral/scatte red

Fig. 6. IU chest X-Ray with assigned tags.

conclusion. Comparison section provides information about a patient's previous medical treatment. The indication shows the symptoms of a disease provided by the patient and patient's metadata. In the findings section, radiologists write their observations (Fig. 5).

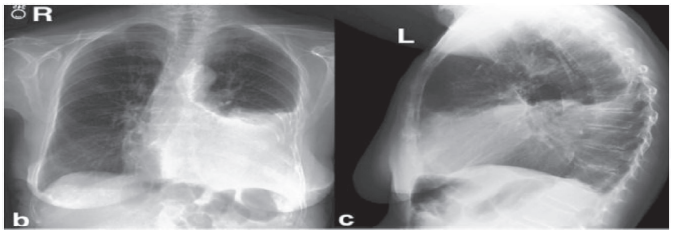
The tags section consists of keywords that contain critical information from the impression and findings section. These tags are the manual encodings of the impression and findings section and are generated using Medical Subject Headings (MeSH) (https://goo.gl/iDvwj2) terms. These sections are also automatic encoded through Medical Text Indexer (MTI) (https://ii.nlm.nih.gov/MTI/) program (Fig. 6).

These tags help in generating the terms for final caption generation. Generated reports using this dataset, by most of the researchers, contains only impression and findings sections [5]. However, the generated description is more detailed. IU Chest X-Ray is provided publicly through Open Access Biomedical Image Search Engine (OpenI) (https://openi.nlm.nih.gov/).

National Institute of Health (NIH) provided a large scale Chest X-Ray14 [4] dataset that consists of 112,120 fronatl chest X-ray images from 30,805 different patients. Each image is assigned one or more lables from 14 thorax disease, or "No Findings". Labels were originally extracted from "Findings" and "Impression" sections of radiologist reports using label extraction tools. This dataset is publically available, however complete diagnostic reports are not presented.

4.4. PEIR gross

The Pathology Education Informational Resource (PEIR) (http://peir.path.uab.edu/library/) digital library provides public access to medical images to be used in medical education. PEIR Gross is



Examination: (Chest PA AND LAT)
Indication: ___ years old woman with ?pleural effusion // pleural effusion
Technique: Chest PA and Lateral
Comparison: ____
Findings: cardiac size cannot be evaluated. Large left pleural effusion is new. Small right effusion is new. The upper lings are clear. Right lower lobe opacities are better seen in prior CT. There is no pneumothorax. There are mild degenerative changes in the thoracic spine.
Impression: Large left Pleural effusion

Fig. 7. Frontal and lateral chest X-Rays with an associated report (MIMIC–CXR).

a sub-collection of PEIR digital library. It consists of 7443 image-caption pairs from different sub-classes of PEIR albums. Images are annotated with tags from the words of the caption with the highest tf-idf score. The first work to use images and corresponding captions from a collection of 21 different PEIR pathology subcategories is Jing et al. [5].

4.5. PadChest

PAthology Detection in Chest radiographs (PadChest) [9] consists of 160,868 chest x-rays from 6 different views associated with 109,931 Spanish reports collected from 69,882 patients of Hospital San Juan de Alicante, Spain. Over 20,000 chest x-rays, labels were extracted manually from reports with the help of expert physicians. For the remaining images, automatic supervised label extraction was done. Overall, 297 labels were extracted that were divided into different categories including 174 labels as different radiology findings, 19 diagnostic labels, and 104 anatomic locations. In our study, we have not found use of PadChest dataset.

4.6. CheXpert and mimic-cxr V2.0.0

CheXpert (Chest eXpert) [7] and Medical Information Mart for Intensive Care-Chest X-ray (MIMIC–CXR) V2.0.0 [8] are recently released publicaly avaialable datasets containing a large number of chest x-rays. In both datasets, annotations are extracted from radiologist reports by using a labeler designed by authors of CheXpert called cheXpert labeler. CheXpert dataset contains 224,316 chest x-ray images collected from 65,240 distinct patients. 14 structured observations are extracted from images where each observation is classified into structured labels as blank, positive, negative or uncertain. This dataset is validated using radiologists' labels (ground-truths) and scores of radiology experts. MIMIC–CXR V2.0.0 is first largest publicaly available chest X-ray dataset containing 377,110 chest radiographs associated with 227,835 diagnostic reports collected from Beth Israel Deaconess Medical Center (BIDMC). The structure of reports (Fig. 7) is the similar to IU Chest X-Ray dataset. Liu et al. [37] is the first to use this dataset for generating medical reports.

Several other datasets exist that are privately owned by researchers such as PACS [23] and CX-CHR [53]. PACS is released by NIH clinical center and contains 216,000 2-dimensional images

with diagnostic reports. CX-CHR datasets consists of 35,500 chest x-ray images with associated Chinese reports.

4.7. Discussion

The datasets discussed in this section differ from each other in terms of number of images, the number of corresponding reports or sentences per image, format of the captions generated, the size of images, and generated captions. PEIR Gross consists of a single sentence caption against each image. ICLEFCaption and BCIDR have single paragraphed reports containing of multiple sentences. IU Chest X-Ray and MIMIC–CXR contain reports having different sections as shown in Fig. 5. Chest X-ray14 has one or more labels rather than complete diagnostics report. All of these datasets have few shortcomings. For example, all datasets are incomplete in the sense that there are no reports against many images. The size of a dataset matters for DL methods, however discussed datasets contain a small number of samples due to number of samples per disease become smaller. This causes model generalization problems. BCIDR, IU Chest X-RAY and PEIR GROSS are of small size (5000, 7470, and 7443 images respectively) which can easily cause over-fitting while using them in state-of-the-art (SoE) DL methods. Although ICLEFCaption dataset is of large size (232,305 images) but this contains large number of noisy images. Similarly, Chest X-Ray14 contains false or ambiguous labels because of the use of automatic label extraction methods. PEIR GROSS contains photographs of clinical incidents rather than images obtained through physical examination. Details of datasets used by reviewed studies are given in Table 1.

5. Evaluation measures

Caption generating methods having complexity in their output are difficult to measure. The evaluation of captions generated by different captioning methods can be intuitively done through an extensive way of human judgment. At the same time, it requires a lot of human effort making evaluation process expensive and difficult to scale up. Also, it suffers from user variances because human judgment is mostly subjective. However, it is also necessary to gage the quality of automatically produced captions in an automatic way using different evaluation measures. These evaluation measures are also used for comparing different caption generating systems based on their capability to generate linguistically equal and semantically correct human-like sentences. These measures are similar to human judgment in varying degrees. The most commonly used evaluation measures for medical image captioning are BiLingual Evaluation Understudy (BLEU) [54], Metric for Evaluation of Translation with Explicit ORdering (METEOR) [55] and Recall-Oriented Understudy for Gisting Evaluation using LCS for longest matching (ROUGE-L) [56]. METEOR, BLEU, and ROUGE are originally evaluation measures of machine translation in which generated sentences and ground truth sentences are compared. Recently new evaluation measures designed especially for general image captioning are Consensus-based Image Description (CIDEr) [57] and Semantic Propositional Image Caption Evaluation (SPICE) [58]. Each evaluation measure is described separately in the next sections.

5.1. BiLingual evaluation understudy (BLEU)

BLEU [54] is most commonly used evaluation measure because it can highly co-relate with human judgments. It is used to find the closeness between the generated and ground truth sentence by matching the different variable-length phrases of generated sentence with the human-written ground truth. Comparison between generated sentence and ground truth is made in n-grams (from 1 to 4-grams). For determining BLEU-1, BLEU-2, BLEU-3, and BLEU-4, both sentences are compared in unigrams, bigrams, trigrams, and 4-grams respectively. For short sentences, a brevity penalty is added to each score.

Geometric mean of all scores multiplied by the penalty score was used in ImageCLEFcaption as an official measure. BLEU measure is suitable only for short captions. BLEU is computed using the following formula [54]:

$$Modified\ Unigram\ Precision\ (p_n) = \frac{Sum\ of\ clipped\ unigrams\ of\ candidate\ caption}{Sum\ of\ all\ unigrams\ of\ all\ groundtruth\ captions}$$

$$Brevity\ Penality\ (BP) = \begin{cases} 1 & if\ c > g \\ e^{(1-g/c)} & if\ c \le g \end{cases}$$

Finally,

$$BLEU\ score = BP * e^{\sum_{n=1}^{N} w_n \log(p_n)} \quad (1)$$

Where c is the total length of candidate caption and g is the total length of effective all ground-truths. p_n is the geometric mean of modified n-gram precisions. N is the total length of n-grams used to calculate $p_n.w_n$ is positive weight between 0 and 1. Range of BLEU score is from 0 to 1.

5.2. Recall oriented understudy for gisting Evaluation- Longest common subsequence (ROUGE-L)

ROUGE-L calculates precision, recall, and F1-measure based on the length of longest common subsequence (LCS) between generated and ground-truth sentences. ROUGE [56] has different variants based on removal of stop word, stemming and n-grams of different length but most commonly used for evaluating a medical caption is ROUGE-L formula given below [56]:

$$Recall\ (R_{lcs}) = \frac{LCS(c,g)}{m}$$

$$Precision(P_{lcs}) = \frac{LCS(c,g)}{n}$$

Finally,

$$ROUGE-L\ score = F_{lcs} = \frac{R_{lcs}\ P_{lcs}\ (1+\beta^2)}{R_{lcs} + P_{lcs}\beta^2} \quad (2)$$

Where $LCS(c,g)$is the length of LCS of candidate caption (c) and ground-truth caption (g). $\beta = \frac{P_{lcs}}{R_{lcs}}$ when $\frac{\partial F_{lcs}}{\partial R_{lcs}} = \frac{\partial F_{lcs}}{\partial P_{lcs}}$.

5.3. Metric for evaluation of translation with explicit ordering (METEOR)

METEOR [55] evaluates the caption by first calculating the BLEU-1 between the generated and human-written ground truth sentence to find the matching results. Then harmonic mean is computed on precision and recall of matching results. Addition to unigrams; stems of sentences, synonyms of words and paraphrase table is also considered for matching both sentences. For longer captions, it considers up to 50% penalty when there are gaps, different ordered words and no-common unigrams matches exist. It can better correlate at the sentence and segment level. METEOR score is computed using the following formula:

$$Unigram\ Precision\ (P) = \frac{u_{cg}}{U_c}$$

$$Unigram\ Recall\ (R) = \frac{u_{cg}}{U_g}$$

$$Harmonic\ mean\ (F\ mean) = \frac{10PR}{R+9P}$$

Table 1
Initial datasets used by different researchers, pre-processing, augmentation done on initial dataset, and splitting of dataset after pre-processing are described. Dash (–) is used if researchers do not mention that detail in their papers.

References	Dataset	Initial Size	Pre-processing	Data augmentation	Training and validation dataset	Test dataset
[24]	IU Chest X-Ray	7470 images Image size = 224 × 224	Image cropping	Yes (X4)	80% for training, 10% for validation	10%
[5]	IU Chest X-Ray PEIR Gross	7470 images 7442 images	Converted all tokens to lowercase, Removed all non-alpha tokens	No	500	500
[38]	Chest X-Ray14	112,120 images	–	–	–	–
[39]	DIARETDB0, DIARETDB1, Messidor	370 images Image size = 1400 × 1152	Cropping, Horizontal flipping	Yes (X 100)	80%	20%
[2]	Ultra-sound scans	4298 images Image size = 300 × 300	Horizontal flipping	Yes	70%	30%
[40]	IU Chest X-Ray	7470 images Image size = 224 × 224	Image scaling, Image normalization, Tokenization on reports	No	80% for training, 10% for validation	10%
[41]	ICLEFcaption	184,614 images with associated captions	divides the training dataset into three subsets of different caption lengths (0–13, 13–30, above 30)	–	164,614 images for training, 10,000 images for validation	10,000 images
[6]	BCIDR	1000 images Image size = 224 × 224	Clip, mirror and rotation operations, Image resize, Image normalization,	Yes	80%	20%
[42]	ICLEFcaption	184,614 images with associated captions	–	–	164,614 images for training, 10,000 images for validation	10,000 images
[43]	IU Chest X-Ray	2775 reports associated with 5550 images Image size = 224 × 224	Image resize, Tokenization on findings and impression section	No	Remaining data	Random 250 reports
[44]	ICLEFcaption	223,307 images with associated captions	Image cropping	Yes	222,314 images	9938 images
[25]	Frontal pelvic X-rays containing hip fractures	50,363 X-rays consisting of 4010 hip fractures	–	Yes	41,032 images with 2923 hip fractures for training, 4754 images with 414 hip fractures for validation	4577 images with 348 hip fractures
[45]	CheXpert	3074 multi-view chest x-rays	Removed samples without multi-view images, concatenated “findings” and “impression” sections, tokenization on reports, error replacement	No	80%	20%
[34]	ICLEFcaption	223,307 images with associated captions	–		222,314 images	9938 images
[33]	ICLEFcaption	184,614 images with associated captions	–	–	164,614 images for training, 10,000 images for validation	10,000 images
[46]	ICLEFcaption	184,614 images with associated captions	Single digits, stop words, special characters, and delimiter removal,	–	164,614 images for training, 10,000 images for validation	10,000 images
[37]	MIMIC-CXR IU Chest X-Ray	327,281 images with 141,783 reports 6471 images with 3336 reports	Tokenization on reports	–	70% for training, 10% for validation	20%
[47]	ICLEFcaption	184,614 images with associated captions	Converted all tokens to lowercase, punctuation and stop-words removal, stemming, Image resizing, flipping, rotation, and cropping	Yes	164,541 images for training, 10,000 images for validation	10,000 images
[48]	ICLEFcaption	70,786 radiology image-concepts pairs Image size = 224 × 224	Concepts filtering, Image normalization, flipping, resizing, mirroring, rotation, and cropping	Yes	56,629 images for training, 14,157 images for validation	–
[49]	IU Chest X-rays	2225 image-reports pairs	Converted all tokens to lowercase, Removed all non-alpha tokens	–	70% for training, 10% for validation	20%
[50]	IU Chest X-Ray CX-CHR	7470 images X-rays from 33,236 patients	Tokenization, Converting all tokens to lowercase, tokens filtering	–	70% for training, 10% for validation	20%
[51]	Radiographic reports from Stanford Hospital	87,127 radiographic reports	Reports tokenization and filtering	–	70% for training, 10% for validation	20%
[52]	ICLEFcaption	Image size = 256 × 256	Converted all captions to lowercase, Image normalization, cropping	Yes	–	–

$$Penality\ (p) = \left(\frac{C}{U_m}\right)^3 * 0.5$$

Finally,

$$METEOR\ score = F\ mean * (1 - P) \tag{3}$$

u_{cg} is the total number of unigrams in candidate caption that are mapped to ground-truth caption. U_c and U_g are total the number of unigrams present in candidate caption and ground-truth caption respectively. C is the total number of chunks (set of unigrams in ground-truth caption that are adjacent to unigrams in mapped candidate caption). $U_m$is the total number of unigrams matched.

5.4. Consensus-based image description evaluation (CIDEr)

CIDEr [57] measure is specially developed for evaluating image caption and is also used for video captioning. It is calculated by finding the cosine similarity between generated caption and ground-truth based on Term Frequency - Inverse Document Frequency (tf-idf) score of each n-gram. Average of scores for n-grams is returned as the final score. Formula of CIDEr is given below:

$$\begin{aligned} CIDEr_n &= \frac{1}{m}\sum_j \frac{g^n(c_i).g^n(g_{ij})}{\|g^n(c_j)\|\|g^n(g_{ij})\|} \\ CIDEr\ score &= \sum_{n=1}^{N} w_n CIDEr_n \end{aligned} \tag{4}$$

Where $g^n(c_i)$ and $g^n(s_{ij})$ are vector forms of tf-idf weight for each n-gram of candidate and ground-truth captions respectively. $w_n$is a uniform weight normally equal to 1 [57].

5.5. Semantic propositional image caption evaluation (SPICE)

Semantic Propositional Image Caption Evaluation (SPICE) [58] measure is based on novel semantic concepts. It extracts objects, attributes, and relationships between them from generated and ground truth captions, and forms a tuple against each caption, containing all above-mentioned information. Precision, recall, and F1-measure is computed to find the final results. SPICE is computed using the following formula [58]:

$$Precision\ (P) = \frac{T(G(c)) \otimes T(G(g))}{T(G(c))}$$

$$Recall\ (R) = \frac{T(G(c)) \otimes T(G(g))}{T(G(g))}$$

Finally,

$$SPICE\ score = F_1 = \frac{2PR}{P+R} \tag{5}$$

T is a function that returns tuples from the scene graph of candidate caption ($G(c)$ and ground-truth caption ($G(g)$. $\otimes$ is a function that returns matching tuples from both graphs.

In this study, we have not found use of SPICE for medical image captioning. BLEU and ROUGE are used more than CIDEr and METEOR. No evaluation measure perfectly matches up with human decisions and they all have strengths and limitations. Although BLEU metric co-relates with human judgments but it performs low in explicit n-gram matching. In BLEU, ROUGE-L, and CIDEr, word ordering affects [59] because exact n-gram matching is used in these metrics, but the ordering doesn't affect SPICE because it performs synonyms matching at sentence level. METEOR also performs stemming, and paraphrase and synonym matching so it performs well at segment or sentence level but semantic similarity suffers. Evaluation score of all measures decreases if the words are replaced with their synonyms [59]. However, it is stated in Elliott et al. [60], that it is impossible to find always perfect correlation between these automatic methods of evaluating image captions and human judgments. Therefore, some researchers Anderson et al. [58] and Kilickaya et al. [59] recommend the use of a joint of captioning measures having various dimensions such as correctness, saliency, thoroughness, and grammaticality.

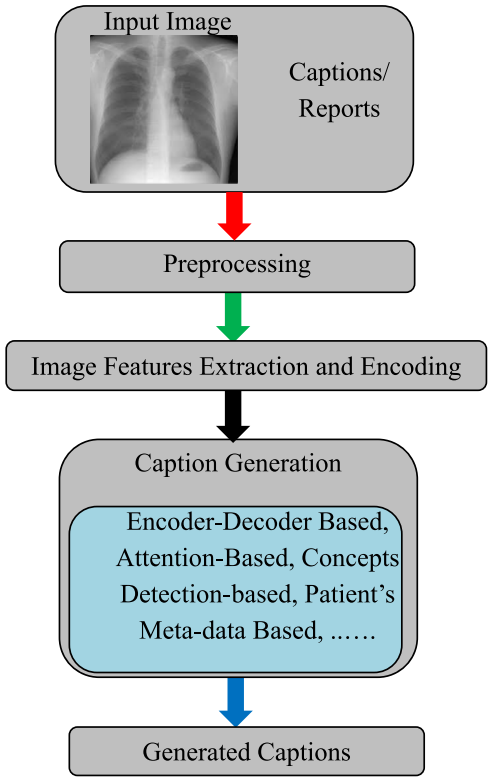


Fig. 8. Deep learning based image captioning.

6. Deep learning based medical image caption generation

Recently natural language like image captioning using deep learning networks [61,62] has gained great success. This motivated the researchers to use deep learning methods for medical image captioning. Most of the existing literature uses encoder-decoder architecture where CNN extracts features from images and encodes them into vector representations of fixed length. These representations are fed into decoder RNN that generates the sequence of words against these vector representations. LSTM or Gated Recurrent Units (GRU), variants of Recurrent Neural Networks (RNN) act as decoder to generate sentences having descriptive semantics and correct grammar. All parameters of these models are trained altogether. Hence, end-to-end training is done in these models such that there is no need to align several independent modules. But, the disadvantage of this approach is that all different types (visual spatial and semantic etc.) of extracted image features are encoded into a single vector representation of fixed length that is used by a decoder as a single input. A general pipeline of DL-based medical image captioning is shown in (Fig. 8).

6.1. Pre-processing

In medical image processing, use of multiple imaging modalities is ubiquitous that helps professionals during the diagnosis process by analysing through different aspects. As an example, interested reader can easily find the use of multi-modalities in disease detection, segmentation, classification, and in generating

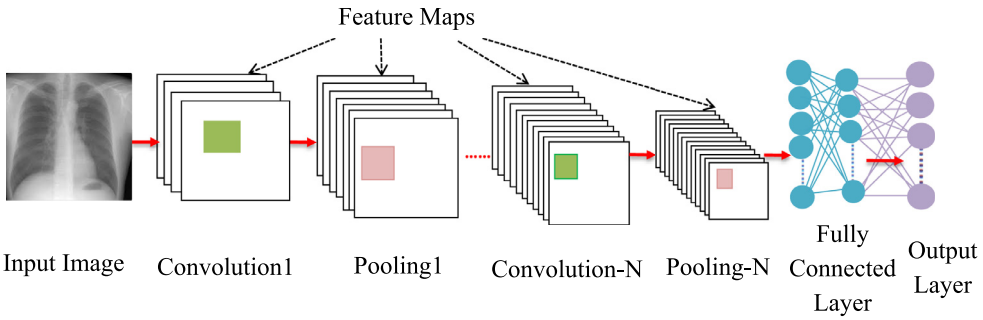


Fig. 9. General architecture of CNN.

descriptions of detected disease. Images are acquired from patients with varying history using different modalities with varying settings.

Acquiring exact information that is free of all type of inadequacies is a challenging task because various types of noise can affect this process. So, there is always need to apply different pre-processing techniques in order to improve its quality. Although medical image pre-processing is a challenging task, but pre-processing enhances the interpretability of images leading to reduction in time complexity and enhancing the accuracy of models. Image pre-processing is also necessary for generating accurate image captions, but may generate worse results if not applied accurately. Using high quality features, models generate captions for input images. Pre-processing methods are applied on both images and captions in the captioning process. The list of such methods may include noise removal from images using different filters, image normalization, mirroring, resizing, cropping, and flipping. The caption part of dataset is also pre-processed for example,token filtering, conversion of all tokens to lowercase, removing all special tokens, stop word removal, applying stemming, delimiters removal, and removing words containing numbers.

Image captioning is a two phase process that involves 1) image features extraction and encoding; and 2) natural language caption generation.

6.2. Phase 1: image features extraction and encoding

Computational models recognize images on the basis of different features (numerical values). Features extraction is a process to extract different features which create machine understanding of image however mostly extracted features are large in number. To reduce the dimensions of data, feature reduction methods are used to filter most pertinent and useful information from an image, and then summarizing it in a reduced set also called a low dimensional feature vector. The desired task is then performed using this feature vector instead of a full-sized input image or high dimensional feature vector. Such techniques help in preventing overfitting, improving accuracy, speeding up training, improving the visualization of data, and enhancing the interpretability of computational models. Extraction of effective features is a critical step that leads to generate more accurate image captions. Deep learning based convolutional neural networks (CNNs) are used for feature extraction and encoding features into vector representation of fixed length to provide high accuracy. CNNs are most popular models for segmentation, detection, and recognition of objects.

CNN consists of an input layer, multiple hidden layers and an output layer. Hidden layers typically consist of convolutional layers, pooling layers, normalization layers and fully connected layer. Increasing number of hidden layers increase the complexity of a model. Each convolutional layer produces feature maps (set of features) which are then fed to next layers. Pooling layers reduces the size of a feature map which help to extract higher level features resulting in hierarchical pipeline of features, i.e. lower-level, mid-level, and high-level features. The result of a pooling layer is fed into a fully connected layer that performs final classification task. General CNN architecture is shown in Fig. 9.

However, DL based CNN relies on a large scale of labeled data (medical images in current context) to achieve remarkable performance on various image analysis tasks i.e. image classification, caption, recognition, segmentation and detection. Since getting a large volume of labeled medical images of high quality is an expensive and difficult task, another technique of supervised machine learning called transfer learning (TL) is required to overcome this challenge. Transfer learning is the process of transferring knowledge from a source domain having a large scale of labeled data to a different but related target domain having a small scale of labeled data. An important characteristic of layered architecture of CNN is that initial convolutional layers extract general features from data which can be reused in other related problems while higher fully connected layers extract specific features depending on the task at hand. Several architectures of CNN exist like VGGNet [12], ResNet [13], InceptionV3 [14], GoogleLeNet [14], AlexNet [63], ZFnet [53], and LeNet [64]. All these models are trained on a large scale of natural images. TL optimizes performance while saving time and allowing us to learn features from natural images and adjust these features or model to adapt the task at hand instead of learning from scratch. In this study, many authors are using different pre-trained (on large scale of natural images) CNN models for feature extraction that are either fine-tuned on domain dataset or used without fine tuning. While transferring a pre-trained model, only last high layer is removed and new predicting layer is added considering the task at hand. In case of fine-tuning pre-trained model, some other higher layers are also replaced along with the last layer. Normally low level layers are remained unchanged because they extract general features and freezing them can save a lot of time spent on training from scratch. During training, weights of the initial layers are not updated and parameters of some high level layers are fine-tuned to adapt the task at hand.

6.3. Phase 2: caption generation

In caption generation step, features extracted from hidden layers of the previous step are briefly described in grammatically and semantically correct natural language sentences using language recurrent model such as GRU [16] and LSTM [17]. Language model predicts the probabilities of the next word based on previously generated word and hidden state until the 'end' token is generated (Fig. 10). The most popular language models are LSTM-based models that overcome the problem of vanishing gradient and generate long sentences. Some of the studies used hierarchical LSTM consisting of sentence-LSTM and word-LSTM. Sentence-LSTM generates topics for sentences given an input

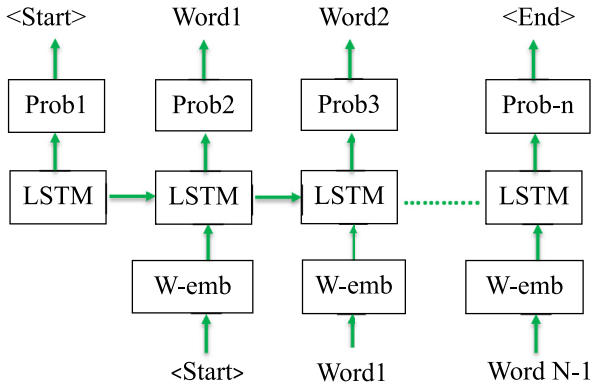


Fig. 10. LSTM network.

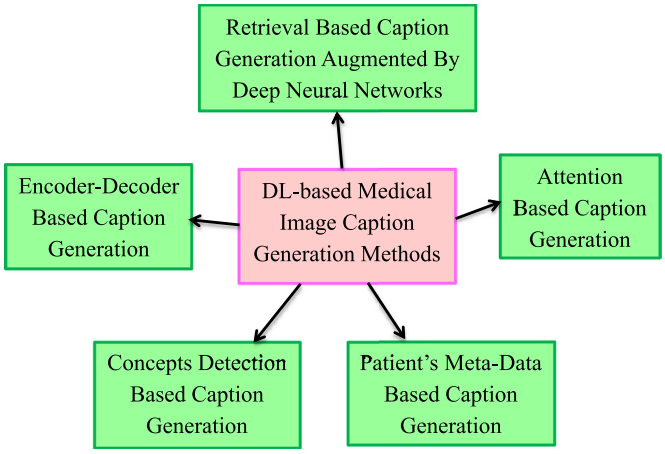


Fig. 11. DL based medical image caption generation methods.

image and word-LSTM generates caption word by word against the topics given by sentence-LSTM. Some of the reviewed work used bi-LSTM in which there are two layers of LSTM.

Focus of this work is on textual tagging and caption generation hence in the following text, we review caption generation methods in detail. Taxonomy of DL-based medical image captioning methods is shown below in Fig. 11.

6.3.1. Encoder-decoder based caption generation

Encoder-Decoder architecture proposed by Kiros et al. [65], is a type of neural networks that is based on neural machine translation in which a sentence is transformed from one language to another, or from features to words. In Machine Translation, a source sentence is given to encoder that encodes the sentence into a vector representation containing the semantic information of an input sentence. This vector is then fed into a decoder that decodes it into a destination sentence. The same idea is applied for caption generation and is proved effective.

An image is given to encoder that extracts features from image and encodes it into a feature vector of fixed length. The feature vector is then fed to a decoder that generates the word by word captions (Fig. 12). CNN can extract rich information from an input image and convert it into embedding vector of fixed length that can be used by many other vision tasks. Therefore, CNN acts as an encoder and RNN (e.g. LSTM, GRU) acts as a decoder for image captioning. Details of reviewed studies using encoder-decoder based caption generation are given in Table 2.

Shin et al. [24] used an encoder-decoder approach first time (to best of our knowledge) using medical images of chest X-rays to not only classify and detect the disease from images but also to generate annotations of images. They first trained Network-In-Network [66] and GoogleNet [14] CNN to classify the images and extracted 17 disease labels corresponding to MESH terms that were being used frequently in the reports while not being co-occurring frequently with other MESH terms and got better performance on GoogleNet. During training, they used mini-batch normalization and dropout techniques to regularize the model. An LSTM or GRU was trained that decoded the CNN embedding of input image to generate the context of detected diseases. In the second training phase, already trained CNN-RNN was re-trained to generate joint an image/text context vectors on image/text pair of domain specific dataset. This time, average pooling was applied on state vectors of RNN resulting better image representations. K-mean clustering was applied on the generated improved image representations that were further used in third training phase where CNN and RNN were trained again on joint image/text context vectors using weights of second training phase. This time CNN classified 57 disease labels and RNN generated context descriptions.

Wu et al. [39] used framework based on Vinyals et al. [61] to generate the description of fundus images. They used a pre-trained CNN having dropout and ensemble techniques for encoding all 37,000 retinopathy images and encoded features were fed into LSTM that decoded the image features into a caption. The generated description told only what abnormalities were found. First they trained model to generate a caption containing four abnormalities (microaneyryms, softexudates, hardexudates, and hemorrhages). They got overall good results but in case of average sensitivity of individual abnormality, results were worst for softexudates because it appeared only in a small part of dataset. Then they experimented by taking softexudates and hardexudates as one abnormality; exudates, and got better results than the previous experiment, but still there was over-fitting due to small scale of dataset.

In Lyndon et al. [47], images were first pre-processed that included rec-scaling, cropping, and applying image augmentation (distorting and random cropping). All associated captions were also pre-processed, including removing the punctuation marks from captions, conversion of captions to lower case, stop-words removal, and applying stemming. After the pre-processing, all captions that originally contained multiple sentences became a single sentence. Pre-processed images were fed into an InceptionV3 CNN that generated image embedding that was further given to LSTM as an initial state only and not used in further time-steps. Sub-sequent to an initial state, at each state LSTM produced output that was further given to a word-embedding layer and then Softmax layer that gave the probabilities of the generated word in the vocabulary. They trained model in two phases. In first phase, weights of CNN were frozen and only LSTM was trained. Then end-to-end training was done on the entire model.

Liang et al. [41] first divided the training dataset into three subsets of different caption lengths (0–13, 13–30, above 30) and presented a method that consisted of three parts. First was a deep model consisting of CNN pre-trained on VGGNet [12] to extract image features and RNN-LSTM to generate a caption. All three subsets of dataset were trained using three different models of this nature. The second part was trained on SVM-classifier that determined the best model to be used. Extracted features were given to SVM classifier that used the three trained models that generated three captions of different lengths. They concluded that training of model on datasets of different caption lengths affects the result and use of SVM along with the CNN-RNN model enhances the performance.

Su et al. [44] used ResNet-152 to extract the features from an input image. Encoded feature vector was fed into LSTM along with a start token to initialize the hidden states of LSTM. At every time step of LSTM, a word was generated based on a hidden state and previously generated word. The feature map was given only in the

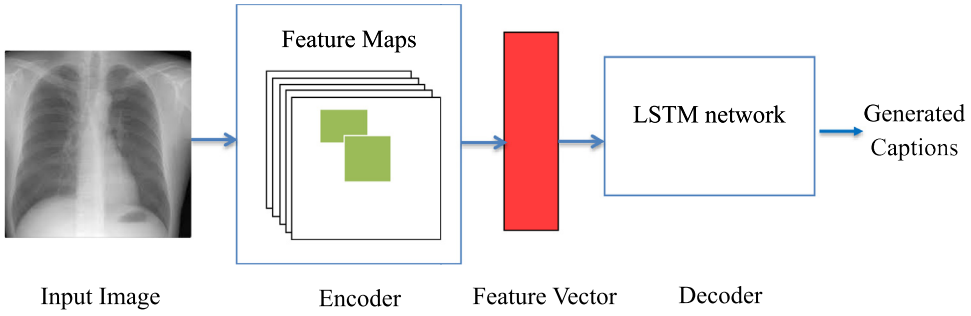


Fig. 12. Encoder-decoder baced caption generation.

Table 2
Encoder Decoder based caption generation. Dash (–) is used if researchers do not mention that detail in their papers.

References	Encoder	Additional Processing	Decoder	Transfer Learning	Regularization
[24]	GoogleNet	Image classification and detection	LSTM, GRU	Yes (fine-tuning)	Dropout, Batch normalization
[2]	VGG-16	Region detection, Classification Regression	LSTM	Yes (fine-tuning)	–
[39]	Batch Normalized CNN [70]	No	LSTM	Yes	Dataset Expansion, Dropout
[47]	Inception-V3	Concepts detection	3- layer LSTM	Yes	Dropout
[41]	VGGNet	SVM classification	LSTM	Yes	–
[46]	Inception-V3	No	LSTM	Yes (fine-tuning)	–
[44]	ResNet-152	Experiment on VGGNet	LSTM	Yes (fine-tuning)	Dataset Expansion
[52]	–	Coding Images into continuous representation	–	–	Early stopping

first time step. They got 0.18 mean BLUE score and ranked second in ImageCLEFFcaption 2018. They also performed experiment on VGGNet encoder using soft attention however they got the best results on ResNet.

Another encoder-decoder based method was designed by Pelka et al. [46] for predicting keywords. The focus of this work was to generate keywords that can be used for image retrieval and classification. They first trained LSTM on an original dataset of image-caption pairs. Then they fine-tuned their CNN-LSTM network using InceptionV3 and Inception-ResNet-v2 and got the best results on InceptionV3.

Caption generation for ultra-sound images by Zeng et al. [2] was divided into two phases: encoder phase and language generation phase, and both of these phases were trained during the training process. In the first phase, a region detection model, Faster-RCNN [67] was trained that classified diseases and detected the focus area in the ultra-sound medical image. It generated the probabilities of a proposed region related to a specific disease and predicted the location of the focus area and encoded into encoding vectors simultaneously that contained less noise and more detailed information of focus area. An image was given to Faster-RCNN that used a CNN to extract features that were further given to the Region proposal Network (RPN) [67] that generated the region proposals in the form of arbitrary feature maps. These feature maps of arbitrary size were converted into fixed size features using Region of Interest (ROI) pooling. Finally, these fixed size feature maps were given to a fully connected layer that further was connected to two layers; a softmax layer to classify diseases and a regressor layer that calculated the offsets. Region proposals with the highest scores were referred as focus areas that were converted into encoding vectors. Then in the second phase, the caption generation model was trained to convert the image feature vectors into descriptions. Encoded vectors were given to a LSTM language generation model that generated the annotations containing the content of focus area.

Spinks et al. [52] first encoded the captions and images into a continuous representation space. Captions were encoded by using adversarially regularized autoenoder (ARAE) [68] through a combination of GAN [69] and the autoencoder and images were encoded using CNN. A Decoder learned on decoding captions generated a caption of a test image by mapping its encoded representation into captions. This model achieved 0.1376 mean BLUE score and was ranked fourth out of five participants in ImageCLEFFcaption 2018.

Discussion: The reviewed research used Transfer Learning (TL) to formulate their NN based algorithms. In Encoder-decoder based architecture, encoder part used a pre-trained (on natural images) model which was then fine-tuned on domain dataset. The researchers normally used single layer LSTM as a decoder except Lyndon et al. [47] in which a three layered LSTM was used. Shin et al. [24], Wu et al. [39], Su et al. [44], and Lyndon et al. [47] used regularization to avoid overfitting. Zeng et al. [2], Pelka et al. [46], and Spinks et al. [52] did not mention any regularization technique for their works.

Captions generated in all these methods were short. Shin et al. [24] achieved good results but generated captions described only the context of classified disease and were simply bag-of-words consisting of 5 words instead of fluent coherent report. Captions generated in Wu et al. [39] was consisting of only abnormalities found in the images. Since Wu et al. [39] focused only on whether the generated caption was related to image or not, evaluation measures used were accuracy, specificity, and sensitivity. Although they used regularization in their model, but still there may be overfitting due to small size of dataset. Although model of Liang et al. [41] showed some enhancement in the performance by training on different models using different sub-datasets with different properties and finally classifying through SVM but, their proposed model was not suitable for generating fully descriptive and complex captions like natural language description. Pelka et al. [46] also generated only keywords instead of a full sentence caption. Zeng et al. [2] used Faster RCNN to perform both tasks; region detection from an image and its corresponding encoding in one model. It gave better results than using two different models separately for detecting and then encoding and also better than captioning models used for ultra-sound images which consider full size images. Their model took less time and used a small number of parameters. However, generation of short descriptions and linguistic and word class prediction errors in caption generation are demerits of this model.

6.3.2. Attention based caption generation

Encoder-decoder architecture is further improved with an attention mechanism. An image contains a lot of information whereas it is unnecessary to describe all that information in an

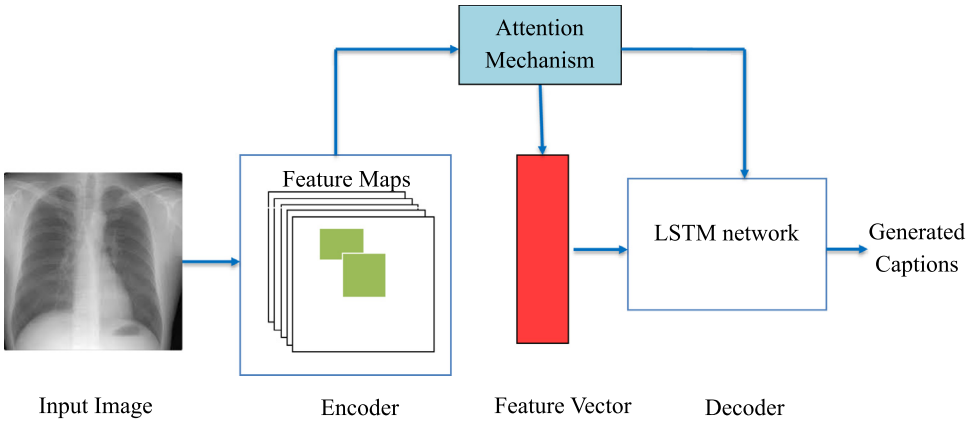


Fig. 13. Attention based caption generation.

image caption. An attention mechanism is introduced in which caption is generated against the only salient information or features of an image by allowing them to come on the forefront. General architecture of this method is shown in Fig. 13.Details of reviewed studies using attention based caption generation are given in Table 3.

The attention mechanism was firstly (to the best of our knowledge) used by Zhang et al. [6] in medical image report generation by proposing their MDNet. They used a pathology dataset of bladder cancer images for this work. Authors stated that subtle changes in bladder can't be accurately identified and discriminated by experienced observers because these image features are suppressed however these semantic features must be used in the generation of efficient diagnostic report. They established direct mapping from cancer images to diagnostic descriptions and used an attention module to perform this modeling more efficiently. They used ResNet [13] as an encoder to extract image features and extended its skip connections to solve vanishing gradient problem. Visual soft attention was computed on extracted features. Attended visual features were given to only initial state of language-LSTM module that generated words at each time step. Since the size of dataset was smaller, they optimized the gradients that behaved like regularization to avoid overfitting. They experimented on fine-tuning pre-trained CNN, pre-trained CNN, and their built network that was jointly trained from scratch. Their jointly trained model outperformed the pre-trained and fine-tuned pre-trained networks.

Jing et al. [5] presented a hierarchal model with a co-attention mechanism in which image was segmented into regions of the same size and given to VGG-19 [12] that extracted visual features from the last convolutional layer. Extracted visual features were given to a multi-label classification (MLC) network that gave the probabilities of tags over pre-determined tag vocabulary. All tags were represented as word-embedding vectors containing high level information that was used as semantic features. Both the semantic and visual features were given to co-attention model that assigned scores to visual and semantic feature vectors and their weighted sum was computed separately that resulted in visual context vectors, semantic context vectors respectively. Both context vectors were concatenated and a joint context vector was produced containing high level, visual and semantic information of an image. In the decoding part, hierarchal LSTM consisting of sentence LSTM and word LSTM was used. At each time step, sentence LSTM produced a topic vector given a joint context vector and updated its state until the control module generates a special "stop" token. Topic vector specified the context of the sentence against which word LSTM produced sentence. Word LSTM took topic vector as its input and produced a sentence word-by-word until control module generated "stop" token. They concluded that their method gives better performance than the models for natural image captioning with visual attention.

Hasan et al. [42] used VGGNet-19 trained on ImageNet as an encoder that extracted the visual features from images from the lower layer of CNN. Extracted visual features were given to a LSTM based decoder that used an attention mechanism and computed soft-attention and gave importance to salient parts of image. At every time step of LSTM, a caption was generated word by word based on context vector containing attended features, hidden states and previously generated word. They got 0.2638 mean BLEU score using originally training images.

Wang et al. [38] presented a method that is similar to Jing et al. [5] but they used ResNet-50 for encoding the input images and RNN based LSTM for generating image captions, but their LSTM was flat instead of hierarchical as used in [5]. Text report was converted into a word embedding and image was given to ResNet-50 that produced feature vector. This vector initialized the hidden states of LSTM and meanwhile multiplied element wise to soft attention. At each time step of LSTM, attended feature, current word and previously generated word was given to LSTM as input that produced output word by word at each time step. The attention score was computed over generated text embedding after which weighted sum was computed to aggregate the representation. The result was multiplied with a hidden state matrix of LSTM and they got final attended text embedding matrix. After that, global max pooling was performed to get the global text embedding sentences. In addition to producing meaningful report using attention, they also performed auto annotation and multi-label classification of image by producing improved attended global visual embedding as spatial regions. In this way, Wang et al. [38] demonstrated that additional information can also be extracted from the hidden states of LSTM.

Liu et al. [37] presented a CNN-RNN-RNN model for generating radiology reports same as Wang et al. [38]. CNN extracts visual features that were further pooled and an average pooled vector was generated which was fed into sentence LSTM to generate topics and an end token that will stop the word LSTM. Topic vector and pooled vector were fed into Word-LSTM that generated the word. An attention was computed over the generated word. At the end, a sentence was generated by concatenating all generated words. The generated report contained duplicate sentences that were removed in post-processing. This model performed slightly better than Wang et al. [38]. They also optimized the model using reinforcement learning to enhance correctness and readability, but still, repetition of sentences was found in the reports. This model

described all negative observations but failed to describe all positive findings.

Focus of Xue et al. [43] was on generating the findings section of IU Chest X-Rays dataset. This task was divided into sub-tasks where each sub-task generated one sentence. They presented a recurrent generation model that maintained coherence and dependency among sentences of the paragraph. In this model, each new sentence was generated by considering the previous sentence and image as joint inputs. They also incorporated an attention mechanism to improve results. An input images pair where each image was from different view, was fed into CNN encoder that was a pre-trained on ResNet-152 that generated global visual features. Global visual feature vector was fed into a sentence generative model that acted as a sentence decoder where a single-layer LSTM predicted the first word of sentence. Whole impression sentence was generated word by word using this sentence LSTM. The recurrent paragraph generative model produced findings paragraph sentence by sentence. It consisted of two parts; sentence encoder and an attentional sentence decoder. Two types of sentence encoders were used in this paper; a Bi-directional LSTM [71] and 1D. Sentence encoder took sentence, generated by sentence generative model as input and generated semantic features of that sentence. Features from different layers were the concatenated form final semantic feature vector. Both semantic features of previously generated sentence and visual features of an image were given to an attentional sentence decoder as a multi-modal input. This recurrent multi-modal model generated the next sentence that was further used while generating the next sentence as a context. This procedure was repeated until an empty sentence was generated to indicate the end of paragraph.

Gale et al. [25] said that reports generated by machine learning algorithms are according to their specified models and have a problem of providing the evidence of their decisions but humans want simple descriptive reports containing the interpretation of findings and decisions for example, where and what features contributed to the conclusive decision. Gale et al. [25] showed that existing radiology report generation methods were inconsistent, did not produce satisfactory results and did not provide decision making process. They proposed a model that produced text descriptions of short length and generated description was simple, meaningful, and explained the decision making process to human in a satisfying way. They argued that training the model on original reports of inconsistent structure and trying to reproduce similar sentences is a difficult task. For simplicity, they trained model on their own created hand labelled descriptive terms that were important in diagnosis and should be in generated report. Their model emphasized on the classification of hip fractures and argued that such minor medical diagnosis should be explained in a diagnostic report that consisted only of two constrained and structured sentence templates; one positive sentence consisting of five placeholders each corresponding to a hand labelled descriptive term and one fixed sentence for negative repose. Model first classified the hip fractures into regions using pre-trained DenseNet and then given to a soft attention module that computed soft attention over each detected region. Attended regions were fed into a recurrent neural network consisting of two LSTM layers and one linear softmax output layer that gave the predictions of the next word in constrained textual sentence.

Yuan et al. [45] said that existing literature used IU-RR small scale datasets like Chest X-Rays due to which performance was not enough good on generated reports. They were the first to solve this problem by pre-training the ResNet-152 encoder on large scale dataset of chest x-ray images, named as CheXpert that gave a robust performance on report generation by learning features related to radiographic. This dataset consisted of multi-view image and reports. It was given to an encoder for pre-training that recognized 14 radiographic observations. Meanwhile, encoder extracted local and global visual features from images. After that another fully connected layer was used that extracted the concepts from training reports. To generate paragraph reports, hierarchical LSTM consisting of sentence LSTM and word LSTM was used. Since visual features contained features related to both frontal and lateral views, these were first given to an attention module that separately computed an attention score over frontal and lateral features and then concatenated them (late fusion). Attended features were fed into sentence LSTM that produced sentence hidden states at each time step. In order to generate a consistent and semantically correct report, extracted concepts were given to an attention mechanism. Attended medical concepts and sentence hidden states were fed into word LSTM that generated the predictions over the next generated word. At each time step of word LSTM, previously generated words were concatenated with the newly generated word. They also performed experiments using early fusion in which first attended features were first concatenated and then attended using hidden states.

Xu et al. [48] used visual attention to generate sequence of medical concepts against medical images. Data augmentation, including random flipping, cropping, and image scaling was performed on training images. First pre-processed images were encoded using pre-trained ResNet-101 and visual features were extracted. Then soft visual attention was computed over features. The decoder used in this paper was simple LSTM that generated a word at each time step based on hidden states, attended context vector, and a word generated in previously time step. The trained model was fine-tuned with dropout and early stopping criteria to avoid over-fitting.

Li et al. [49] first classified and localized the disease to annotate the X-ray images and then generated sentences corresponding to disease to form a report. They used DenseNet-121 to annotate the image as either with MeSH terms as a disease label or with “Normal” label. They localized the image by extracting a region of interest using a bounding box. They used thresholding to localize that region. If the image was labelled as “Normal”, a normal report was generated using attention based LSTM and if the disease region was cropped then the corresponding report was containing both abnormalities from the cropped image and normalities from the original image. They used ResNet-121 as an encoder that generated feature map against either the cropped image or original image. Feature maps were fed into visual attention based LSTM to generate sentences. They used only frontal images from IU X-Rays dataset.

Discussion: In our review, we found that research community is using transfer learning, mostly, to build their captioning models as discussed in the last section. In these models, pre-trained (on natural images) encoders that were either fine-tuned on domain dataset or used without fine tuning, except Zhang et al. [6] in which authors designed their own encoder and trained from scratch. Diagnostic reports generated by Zhang et al. [6] described only five types of bladder features based on the appearance of cells (Fig. 4) that was a less complex problem than to generate a complete radiology diagnostic report. Papers classified under attention based caption generation category, computed the attention over only visual spatial features except Jing et al. [5] in which authors also calculated semantic attention over extracted features. Most of the studies, in this section, improved results by using hierarchical LSTM as a decoder that consisted of sentence-LSTM and word-LSTM. Although Jing et al. [5] achieved good performance due to hierarchical LSTM, but their generated reports contained repetitions of words. Ignoring the contextual coherence in their hierarchical model may be the reason for these repetitions. Wang et al. [38] experimented on OpenI dataset, similar to Jing et al. [5], but didn’t provide final evaluation results. Focusing on their experi-

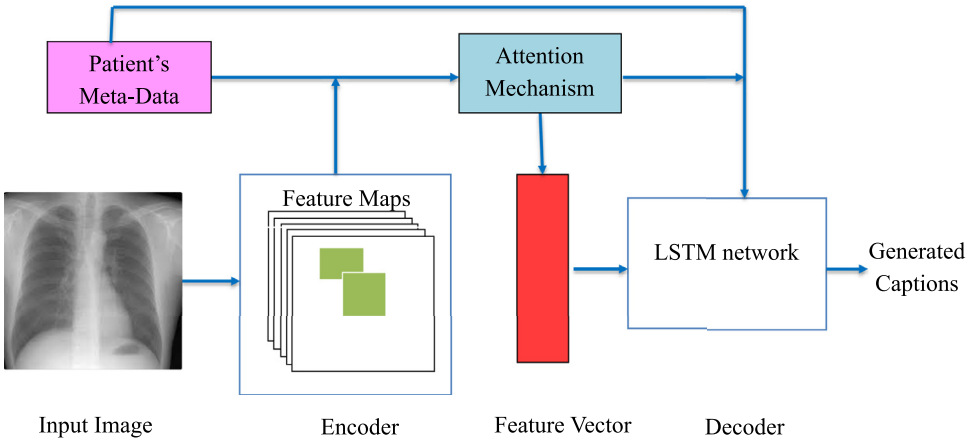


Fig. 14. Patient's meta-data baced caption generation.

ment on Chest X-Rays dataset and viewing the generated reports, it is clear that results were worse than OpenI results of Jing et al. [5]. This may be due to the use of flat LSTM for decoding the textual reports. Although generated reports by Xue et al. [43] were coherent but still had fewer repetitions in reports. Moreover, there were missed abnormalities in some generated reports. The cause of this erroneous behavior may be the training of the model on a small dataset containing a small number of abnormal samples. Sentences that were not present in ground truths were not generated very well. This can also be due to difficulty in learning syntactic correctness in small datasets. Gale et al. [25] could not accurately identify the location of fracture, however described the nature of fracture in simple sentences generated by their model. The results of Yuan et al. [45] showed that incorporating fusion mechanism and concepts information in the decoding process, improves the performance where late fusion gave better results than early fusion. Training encoder on domain specific dataset yielded better results. But due to small scale dataset, their model was not better in generating unseen sentences.

6.3.3. Patient's meta-data based caption generation

A standard medical report usually contains an important section named as background in which critical information such as a patient's metadata, disease symptoms and information about a patient's previous treatments are discussed. A simple way to add this information in caption generation procedure is to concatenate background section with findings section and use a concatenated single text for training model. But, this requires the sufficient training of model to distinguish them. To avoid this overhead, patient's background information is encoded separately and this encoded representation guides the decoder to generate more accurate reports. A simple structure of this method is shown in Fig. 14.

Zhang et al. [51] was the first work (to the best of our knowledge) that used background information in generating an impression statement of a medical reports. The authors presented a customized model for summarizing radiological findings into impression statement by encoding both findings and background section of a radiology report. The proposed model was based on two existing frameworks; neural sequence-to- sequence model and a pointer generator network. Both findings and background sentences were encoded separately using bi-LSTM and then attention was computed over both encoded information. During the decoding phase, encoded findings and background was fed into LSTM that generated a word at each time step using the decoded background information and previously generated word. Role of the pointer generator network was that model was allowed to copy the word directly from the finding section. For this purpose, again attention was computed over the word to be copied.

Multi-attention mechanism presented by Huang et al. [40] consisted of spatial and channel attention. Channel attention told which or what entity is this and spatial attention informed about the spatial location of entity. A hierarchical RNN was used to generate paragraphs. Meanwhile, a patient's background information was fused with word-embedding. An image was given to ResNet encoder that extracted the feature maps. Input feature maps were converted into their channel vectors on which mean pooling was applied and they got average channel vectors on which weighted channel attention was computed. For spatial attention, input feature maps were first flattened into 1D vector where each value was the position of visual feature on the feature map. Flattened vector was given to a single layered neural network that gave spatial attention after that element wise multiplication was done on the output of both attentions. Attended feature maps were fed into a decoder that consisted of three parts; sentence RNN, background information fusion module, and word-RNN. Sentence RNN took the attended feature vectors and generated two types of information; a topic vector against which sentence to be generated and a probability to determine whether to stop or continue the sentence LSTM. If the probability was more than some specified threshold, sentence LSTM stopped, and word LSTM also stopped. Embedded background information was given to bi-LSTM whose output was hidden states that were given to a multi-layer perceptron attention mechanism that calculated the alignment score. These were normalized into vector using softmax. This attended background embedding was added to the report's word-embedding that gave a new fused word embedding. Word-RNN was bi-LSTM that initially took topic vectors and special word ⟨start⟩ as first and second input and subsequent inputs were the word embedding fussed with background information. At each time step, the last hidden state was used to get the word's prediction over vocabulary and ⟨end⟩ token to indicate the end of a sentence. After every word generated, it was concatenated to form the sentence. They concluded that their method gives better performance than the models for natural image captioning because of the attention mechanism and adding background information.

Discussion: Little work is done in which background information is incorporated for generating reports and achieved good results. Zhang et al. [51] used transfer learning and regularization in their model but they gave findings paragraph as an input instead of a medical image and generated only an impression sentence. Huang et al. [40] was the only one who computed spatial and channel attention rather than only visual attention. But, due to small dataset, there was significantly difference between the gen-

Table 3
Attention based caption generation. Dash (–) is used if researchers do not mention that detail in their papers.

References	Encoder	Additional Processing	Type of Attention	Decoder	Transfer Learning	Regularization
[5]	VGG-19	Multi-label Classification to predict tags	Visual spatial and semantic	Hierarchical LSTM (Sentence,Word)	Yes (fine-tuning)	Yes (Early Stopping)
[6]	Own designed new ResNet	Symptom description based image retrieval	Visual spatial	LSTM	No	Yes (Gradient Optimization)
[43]	ResNet-152	–	Visual attention over an image and semantic attention over sentence	Hierarchical BiLSTM	Yes (Pre-trained)	–
[25]	DenseNet	Manual labeling of images	Visual spatial	BiLSTM	Yes (Pre-trained)	Yes (Dropout, augmentation)
[38]	ResNet-50	Auto-annotation and classification of images	Visual spatial	LSTM	Yes (Pre-trained, fine-tuning)	Yes (Dropout, L2 regularization)
[45]	ResNet-152	Classification of chest radiographic observations	Visual spatial	Hierarchical LSTM (Sentence,Word)	Yes (pre-trained on CheXpert dataset)	–
[37]	–	Disease classification	Visual spatial	Hierarchical LSTM (Sentence,Word)	–	–
[42]	VGGnet-19	Concepts generation	Visual spatial	LSTM	Yes (fine-tuning)	Yes (Dropout)
[48]	ResNet-101	Concepts detection	Visual spatial	LSTM	Yes (Pre-trained)	Yes (Dropout, early stopping)
[49]	RseNet-151	Disease classification, localization	Visual spatial	LSTM	Yes (Pre-trained)	–

Table 4
Patient's meta-data based caption generation. Dash (–) is used if researchers do not mention that detail in their papers.

References	Encoder	Additional Processing	Decoder	Transfer Learning	Regularization
[40]	ResNet-152	Spatial and Channel Attention	Hierarchical BiLSTM (Word,Background, Sentence LSTM)	Yes	Yes (Dropout)
[51]	BiLSTM	Attention mechanism	LSTM	No	–

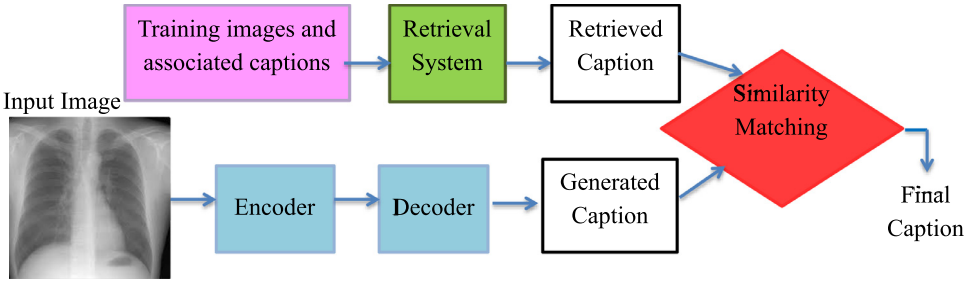


Fig. 15. Retrieval baced caption generation augmented by deep neural networks.

erated and the original report. Details of reviewed studies using a patient's meta-data based caption generation are given in Table 4.

6.3.4. Retrieval based caption generation augmented by deep neural networks

One approach of generating image captions is retrieval-based in which new caption is produced for the input image by retrieving similar caption or set of captions from an existing database. New produced caption can either be the most similar top retrieved caption or concatenation/aggregation of all retrieved captions. Although generated caption is fluent and correct in grammar but a disadvantage of this approach is that captions against already existing image features cannot adapt to novel objects and scenes resulting in an irrelevant caption. Encouraged by a wide use of deep neural networks in caption generation, researchers are combining them in retrieval based caption generation to increase the capability of retrieval based captioning to describe new images correctly. A general structure of this method is shown in Fig. 15 in which caption generated through a deep neural network is compared with the retrieved caption through any retrieval system. The caption having maximum score is associated as a final caption to the input image.

Liang et al. [41] first divided the training dataset into three subsets of different caption lengths (0–13, 13–30, above 30) and presented a method that consisted of three parts. First was a deep model consisting of CNN which was pre-trained on VGGNet [12] to extract image features and RNN-LSTM was used to generate captions. All three subsets of dataset were trained using three different models. The second part was training an SVM-classifier. Extracted image features were given to SVM classifier that used the three trained models and generated the captions of three lengths corresponding to three models. This part determined the best model to be used and then that model generated the caption. The third part of the proposed model was caption retrieval that applied the nearest neighbor to retrieve a similar image and its caption. The retrieved caption was aggregated with the predicted caption if the Euclidean distance between CNN features of the retrieved image and CNN features of the input image was greater than some specified threshold. The authors experimented with normalized and non-normalized features. They got the best result with 2600 mean Blue score on normalized features.

Ben abacha et al. [33] predicted UMLS concept unique identifiers (CUI's) from medical images to generate captions. To detect concepts, authors used information retrieval based approach, in which every training image was given to open-I as query. In response, 10 results with captions were retrieved and the most similar caption was annotated to get concept's unique identifiers (CUI). Open-I retrieval-based approach gave the best results outperforming all runs with 0.1718 mean F1-score. For caption prediction task, for each image, similar images were retrieved from open-

Table 5
Retrieval based caption generation augmented by deep neural networks. Dash (–) is used if researchers do not mention that detail in their papers.

References	Encoder	Additional Processing	Decoder	Transfer Learning	Regularization
[33]	GoogleNet for multi-label classification	Concepts prediction	–	Yes (fine-tuning)	–
[41]	VGGNet	SVM classification	LSTM	Yes	–
[50]	DenseNet	Reinforcement learning, attention mechanism	Hierarchical RNN	Yes (fine-tuning, pre-trained)	–
[34]	Inception-V3	LIRE retrieval System	–	Yes (pre-trained)	–

I along with their captions. The caption of top scoring retrieved image was associated to the input image with 0.5634 mean BLUE score.

Zhang et al. [34] used retrieval-based approach for caption prediction and got the highest score in this sub-task of ICLEFcaption edition 2018. They used LIRE (Lucene Image Retrieval) in which training images were indexed, according to color and texture features and retrieved the most visually similar images of test image from training images, selected candidate captions of top 3 similar images and combined candidate captions to generate a new caption with 0.2501 mean BLUE score. They also combined the top 3 detected CUI (Concept Unique Identifier) of the retrieved similar images of test image using CNN and LDA (Latent Dirichlet Allocation), and combined the UMLS terms of each CUI with the already generated caption with 0.2343 mean BLUE score.

A hierarchical procedure that decides whether to generate a new caption or retrieve the caption from a template corpus was used by Li et al. [50]. Input images were given to encoder that generated the context vector which was fed into a stacked RNN module that generated sentence topics. Each sentence topic was fed into a retrieval module that generated the probabilities over the task of generating a new caption and retrieving from off-the-shelf template corpus. A new caption generation module was activated if the newly generated caption had greater probability. Both retrieval module (that decides whether to retrieve or generate a new caption) and new caption generation module (that generates new caption from scratch) were updated by some reinforcement algorithm in which sentence-level and word- level rewards were used against these two modules respectively. The caption generation module generated new caption based on context vector and hidden states of RNN. Retrieval approach got higher performance than generating new caption.

Discussion: Although better performance can be achieved combining retrieval-based captioning with deep neural network but still a little work is done in generating medical reports or captions of medical images. In Liang et al. [41], although training of a model on datasets of different caption lengths affected the results, use of SVM along with CNN-RNN model enhanced the performance and using the retrieved caption can give better results. But, the proposed model was not suitable for generating fully descriptive and complex captions. In approach proposed by Li et al. [50], caption retrieved related to a sentence topic generated by RNN was better than generating new caption using encoder-decoder architecture, however retrieving a caption from template corpus lacked in describing some abnormal findings. On the other hand, caption generation module generated caption containing abnormal findings in high precision. They all did not mention any regularization in their papers. Details of reviewed studies using this method of caption generation are summarized in Table 5.

6.3.5. Concepts detection based caption generation

Main idea of this method is to first generate medical concepts from the visual appearance of medical images and then use that generated concepts as a caption. These concepts are considered as individual elements from which captions are generated. A simple structure of concepts-detection based caption generation is shown below in Fig. 16.

UMLS concept unique identifiers (CUI's) from medical images were predicted by Ben abacha et al. [33] to generate captions. They divided dataset of 164,614 images associated with 20,463 CUI's into two subsets. The first subset contained 92 unique CUI's corresponding to at least 1500 samples of training data. Second subset consisted of 239 unique CUI's associated with at least 400 training samples. To detect concepts, they trained GoogleNet on random 200 samples from each subset to perform multi-label classification. Then associated UMLS-terms, groups, and semantic types of detected CUI's were generated, and grouped as a caption. This approach gave the best results with 0.2247 mean BLEU score among all submitted runs.

An encoder-decoder framework was used by Hasan et al. [42]. VGGNet-19 trained on ImageNet was used as encoder that extracted the visual features from images fed by lower layers of CNN. Extracted visual features were given to a LSTM based decoder that used an attention mechanism and computed soft-attention and gave importance to salient parts of image. At every time step of LSTM, CUI was generated by training concept detection dataset. They replaced the generated CUI's with their associated longest medical term as a caption and got 0.1801 mean BLEU score. Then they also experimented on using all medical terms associated with detected CUI's as a caption and got better results than replacing with only longest medical term as a caption. They got 0.3211 mean BLEU score. On the other side, generated captions using all medical terms gave better results, but there was no coherence in the generated caption.

Discussion: Like other categories, we have found little work on generating medical captions using this method. Most of the work used transfer learning and only Hasan et al. [42] added an attention mechanism in their model. This may be the reason of their higher score than Ben abacha et al. [33]. These both papers were the part of ICLEFcaption2017 competition. They used the same dataset but did not provide any dataset pre-processing details in their papers. Hasan et al. [42] used dropout as a regularization technique whereas Ben abacha et al. [33] did not mention any regularization. On the down side, captions generated by both authors are single sentenced rather than a complete medical report. Details of reviewed studies using concepts detection based caption generation are summarized in Table 6.

7. Comparison of state-of-the-art captioning methods

While no experiments are done for formal evaluation, we offer an analysis of the results and performance as reported in different methods reviewed in this study. Different techniques of DL based captioning methods are compared on datasets using evaluation measures in Table 7. In encoder-decoder based approach, Shin et al. [24] achieved a higher BLEU score on IU Chest X-Ray

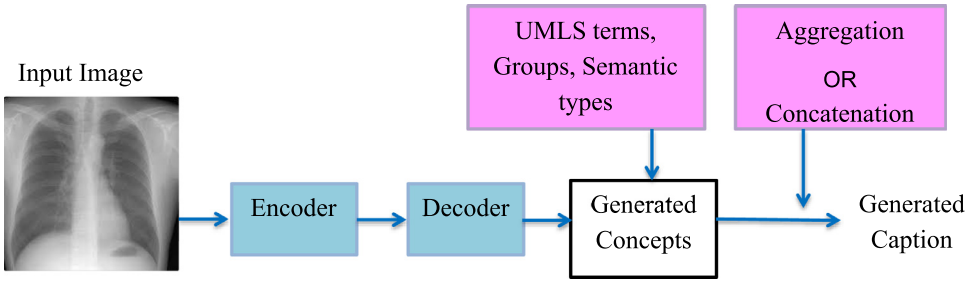


Fig. 16. Concepts detection based caption generation.

Table 6
Concepts detection based caption generation augmented by deep neural networks. Dash (–) is used if researchers do not mention that detail in their papers.

References	Encoder	Additional Processing	Decoder	Transfer Learning	Regularization
[33]	GoogleNet	Concepts detection	–	Yes (fine-tuning)	–
[42]	VGGNet-19	Visual soft attention	LSTM	Yes (fine-tuning)	Yes (Dropout)

Table 7
Comparison of state-of-the-art medical image captioning methods on datasets using evaluation measures. Dash (–) is used if researchers do not use that measure.

Method	Datasets	References	BLEU1	BLEU 2	BLEU 3	BLEU 4	ROUGE	METEOR	CIDEr	Mean BLEU
Encoder-Decoder based	IU Chest X-Ray	[24]	**0.793**	**0.091**	0.0	0.0	–	–	–	–
	ICLEFcaption	[47]	–	–	–	–	–	–	–	0.0982
		[41]	0.13	0.06	0.02	0.01	0.113	0.043	0.053	
		[46]	4	1	6	2	–	–	–	–
		[44]	–	–	–	–	–	–	–	0.0749
		[52]	–	–	–	–	–	–	–	0.1800
			–	–	–	–				0.1376
Attention based	IU	[5]	0.517	**0.386**	0.306	0.247	0.447	0.217	0.327	–
	Chest	[43]	0.464	0.358	0.270	0.195	0.366	0.274	–	–
	X-	[37]	0.369	0.246	0.171	0.115	0.359	–	1.490	–
	Ray	[49]	0.419	0.280	0.201	0.150	0.371	–	0.553	–
	ICLEFcaption	[42]	–	–	–	–	–	–	–	0.2638
		[48]	–	–	–	–	–	–	–	0.2316
	PEIR Gross	[5]	0.300	0.218	0.165	0.113	0.279	0.149	0.329	**–**
	BCIDR	[6]	91.2	82.9	75.0	67.7	70.1	39.6	2.04	**–**
	Chest X-Ray14	[38]	0.2860	0.1597	0.1038	0.0736	0.2263	0.1076	–	**–**
	CheXpert	[45]	**0.529**	0.372	**0.315**	**0.255**	**0.453**	**0.343**	–	**–**
	MIMIC-CXR	[37]	0.352	0.223	0.153	0.104	0.307	**–**	1.153	**–**
Patient's meta-data based + Attention	IU Chest X-Ray	[40]	0.476	0.340	0.238	**–**	0.347	**–**	0.297	**–**
Retrieval based	ICLEFcaption	[33]	–	–	–	–	–	–	–	**0.5634**
		[41]	–	–	–	–	–	–	–	0.2600
		[34]	–	–	–	–	–	–	–	0.2501
Retrieval based + Attention	IU Chest X-Ray	[50]	0.438	0.298	0.208	0.151	0.322	–	0.343	–
	CX-CHR	[50]	**0.673**	**0.587**	**0.530**	**0.486**	**0.612**	–	**2.895**	–
Concepts detection based	ICLEFcaption	[33]	–	–	–	–	–	–	–	0.2247
Concepts detection based + Attention	ICLEFcaption	[42]	–	–	–	–	–	–	–	**0.3211**

dataset than other models applying the same method on ICLEFcaption. Results on IU Chest X-Ray were further outperformed in Jing et al. [5] who used hierarchical LSTM and attention based method in which only important features were focused for captioning. In attention based approach, Yuan et al. [45] is superior to all other models under this method and gave the highest score on CheXpert dataset. Incorporating a patient's metadata (background) in the attention mechanism also got better results on IU Chest X-Ray in Huang et al. [40] than models of Xue et al. [43], Liu et al. [37] and Li et al. [49] in which no background information was added. In Hasan et al. [42], concepts detection based method which also applied an attention mechanism is used and they got higher Mean BLEU score on ICLEFcaption dataset than simple encoder-decoder based approach [33]. Retrieval based method in Zhang et al. [34] performed better than the encoder-decoder based approach on ICLEFcaption dataset. Performance is further slightly improved in Liang et al. [41] due to training model on datasets of different caption lengths, using SVM along with a CNN-RNN model and using the retrieved caption. But, the proposed model was not suitable for generating fully descriptive and complex captions. Highest Mean BLEU score on ICLEFcaption datasets is achieved in Ben abacha et al. [33] using a retrieval based method. The retrieval based method augmented by an attention mechanism [50] outperformed all discussed captioning methods and achieved the highest score on CX-CHR dataset.

Scores in Table 7 show that performance of a captioning depends on dataset size, parameters of underlying model and evaluation measures and vary accordingly. So, comparing different models of different structures is a hard task. It is found that the attention based method is most widely used and gives better performance either it is being applied simply on an encoder-decoder based method or used on a patient's metadata or detected concepts to be used for generating caption.

8. Findings, limitations and potential future directions

Writing medical report in textual form may be error-prone, time-consuming, and laborious, whereas medical professionals have to examine many medical images per day. However, an automatic report generation using AI techniques can give better results.

Researchers are presenting different deep-learning based models for this purpose among which attention-based caption generation is considered better. Different imaging modalities are being used in medical such as X-rays, CT-Scans, magnetic resonance imaging (MRI), ultra-sound, and Positron Emission Tomography (PET). It is necessary to understand imaging modalities before generating corresponding medical reports so that relevant information can be extracted. Our findings after reviewing existing literature are listed below:

- Most of the researchers used medical domain datasets by fine-tuning them on encoders that are pre-trained on natural image datasets except Zhang et al. [6] in which they designed their own encoder and trained from scratch on medical images.
- Dataset that is mostly used in reviewed studies is IU-Chest X-rays that can be their comparison parameter.
- Most of the papers discussed their dataset pre-processing details and only few of them did not mention any pre-processing done in their implementation.
- Different pre-trained encoders are used to extract features from medical images and mostly extracted features are based on visual appearance.
- Attention was applied on only visual features in most of the studies, however few researchers computed semantic and channel attention along with visual attention.
- Most of the studies were ignoring spatial and channel attention.
- Retrieval-based captioning improved by a neural network produced very good results whereas it failed in generating complex captions.
- Captions generated by all the researchers were either single sentence or in a paragraph rather than a complete and structured report containing of different sections. Only Xue et al. [43] generated impression statement also along with findings paragraph in their report.
- Mostly used regularization techniques in reviewed literature were data augmentation and dropout. Some of them used early stopping and L2 regularization.
- Most of the studies were using transfer-learning with fine-tuning the encoder. Some of them also performed additional experiments using pre-trained encoder. But, training the model from scratch using domain dataset can improve the performance.
- All studies encoded the whole input image except Zeng et al. [2] in which they first detected the abnormal regions and then encoded only that region against which captions were generated. This is a good approach that reduces the time complexity because of a small number of parameters in the model.
- The caption generated using hierarchical LSTM gave better performance than using flat LSTM. But, ignoring contextual coherence caused repetitions in the generated report.
- There was not a single paper in which a complete medical report consisting of different sections is generated.

DL based methods has made great progress in generating captions against radiology images. Different medical image captioning techniques have been proposed by the research community. But these methods are not applicable to real world applications because of various challenges in existing models that should be addressed and still, there is much room to increase the performance of generating radiology reports. Some limitations of the existing methods and potential future research directions to fill the gap between existing academic researches and the real world radiology report generation are given below:

- Existing publicly available datasets for medical image captioning are limited in number and there is need to generate more large size datasets.
- All publicly available medical datasets for medical image captioning are imbalanced in the sense that there are no reports against many images.
- All publicly available datasets for medical image captioning contain a small number of abnormal samples where the number of samples per disease becomes even smaller. This causes model generalization problems.
- The size of dataset affects the training and performance of a computational model. Due to a small sized dataset, usually NN model suffers from overfitting, and also there may be a difference between the generated and original captions or reports.
- Due to small dataset, learning syntactic correctness is difficult resulting missed information in generated captions/reports.
- Most widely used evaluation measures are not originally coined for bio-medical image captioning. There is need to discuss how these are appropriate for measuring medical image captioning.
- Current methods are not sufficient enough to produce a coherent, grammatically and semantically correct and complete radiology report. There is need of more strong language modeling structures to monitor long dependencies.
- Adding spatial attention may better guide the decoder to generate captions.
- Performance of encoder may be improved on extracting features by using channel attention along with visual attention.
- Evaluation measures specific for medical image captioning are needed because BLEU, ROUGE and CIDEr do not consider the word ordering and give 100% score even though ground-truth caption shows normal finding but model generated caption includes abnormalities. Similarly, these measures do not capture negation and punctuation in reports accurately.
- Incorporating background information of a patient can enhance the performance of captioning model.
- Generating large scale balanced medical image datasets for radiology reporting and removing noisy and scattered information from existing datasets are ongoing challenges to improve the quantity and quality of datasets.
- Existing datasets do not have clinically annotated images having bounding boxes around normal and abnormal regions of image. But, using object detection models to focus only on an abnormal part of the image can improve the performance of a CNN model on learning relationships among features and extracting only fine-grained features.
- Most of the existing work has focused only on few anatomical parts of body (i.e. Chest X-rays) while ignoring other fatal diseased parts like breast and brain, etc. and no corresponding imaging dataset is available for captioning.
- For 3D imaging data i.e. brain MRI dataset, no work has been done on image captioning.
- Generated reports can be extended to medical images retrieval purposes.
- Building multipurpose radiology reporting models that can detect multiple diseases simultaneously is an important research direction.

9. Conclusion and future work

In this work, we attempted to provide an organized reference for people attracted to medical report generation of medical images using deep learning. We tried to review every recently published research in this problem domain thoroughly and presented it in a detailed and structured manner. We observed that a lot of work is done using a simple encoder-decoder based framework whereas attention-based captioning is being used widely. Moreover, different emerging models for medical image captioning have prospective practical uses in real world medical sectors. The medical sector can gain great benefit from these automatic medical re-

port generation systems. It is hoped that this all work will be actually applied in medical sectors in the near future. This will reduce human efforts and time spent on manual examination and writing medical reports and also generated reports will have high accuracy and precision.

On the other hand, more work is needed to be done such as there is not a single model that can generate complete structured medical report (similar to natural language description) having different sections that are normally part of hand-written medical reports. Different modalities of medical images are being used, better algorithms for generating reports using these modalities are needed to be analysed and tested.

As our own future work, we intend to develop our own large size dataset in consultation with domain experts. We also intend to build industry standard automatic report generation system that will generate a complete medical report according to the needs of radiologists which is missing in the existing literature.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.patcog.2021.107856.

References

[1] A. Brady, R.Ó. Laoide, P. McCarthy, R. McDermott, Discrepancy and error in radiology: concepts, causes and consequences, Ulster Med. J. 81 (2012) 3–9.

[2] X. Zeng, L. Wen, B. Liu, X. Qi, Deep learning for ultrasound image caption generation based on object detection, Neurocomputing (2019), doi:10.1016/j.neucom.2018.11.114.

[3] D. Demner-Fushman, M.D. Kohli, M.B. Rosenman, S.E. Shooshan, L. Rodriguez, S. Antani, G.R. Thoma, C.J. McDonald, Preparing a collection of radiology examinations for distribution and retrieval, J. Am. Med. Informatics Assoc. 23 (2016) 304–310, doi:10.1093/jamia/ocv080.

[4] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R.M. Summers, ChestX-ray8 : hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases, (2017). https://doi.org/10.1109/CVPR.2017.369.

[5] B. Jing, P. Xie, E.P. Xing, On the automatic generation of medical imaging reports, ACL 2018 - 56th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap 1 (2018) 2577–2586, doi:10.18653/v1/p18-1240.

[6] Z. Zhang, Y. Xie, F. Xing, M. McGough, L. Yang, MDNet: a semantically and visually interpretable medical image diagnosis network, in: Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017. 2017-Janua, 2017, pp. 3549–3557, doi:10.1109/CVPR.2017.378.

[7] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D.A. Mong, S.S. Halabi, J.K. Sandberg, R. Jones, D.B. Larson, C.P. Langlotz, B.N. Patel, M.P. Lungren, A.Y. Ng, CheXpert: a Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison, Proc. AAAI Conf, Artif. Intell. 33 (2019) 590–597, doi:10.1609/aaai.v33i01.3301590.

[8] A.E.W. Johnson, T.J. Pollard, S.J. Berkowitz, N.R. Greenbaum, M.P. Lungren, C.Y. Deng, R.G. Mark, S. Horng, MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports, Sci. Data. 6 (2019) 317, doi:10.1038/s41597-019-0322-0.

[9] A. Bustos, A. Pertusa, J.-.M. Salinas, M. de la Iglesia-Vayá, PadChest: a large chest x-ray image dataset with multi-label annotated reports, (2019). http://arxiv.org/abs/1901.07441.

[10] C. Eickhoff, I. Schwall, A.G.S. De Herrera, H. Müller, Overview of imageclefcaption 2017 - Image caption prediction and concept detection for biomedical images, CEUR Workshop Proc (2017) 1866.

[11] A.G. Seco De Herrera, C. Eickhoff, V. Andrearczyk, H. Müller, Overview of the ImageCLEF 2018 caption prediction tasks, in: CEUR Workshop Proc., 2018, p. 2125.

[12] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc., 2015, pp. 1–14.

[13] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2016-Decem, 2016, pp. 770–778, doi:10.1109/CVPR.2016.90.

[14] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 07-12-June, 2015, pp. 1–9, doi:10.1109/CVPR.2015.7298594.

[15] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling, (2014) 1–9. http://arxiv.org/abs/1412.3555.

[16] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, in: EMNLP 2014 - 2014 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf., 2014, pp. 1724–1734, doi:10.3115/v1/d14-1179.

[17] S. Hochreiter, J. Schmidhuber, Long Short-Term Memory, Neural Comput 9 (1997) 1735–1780, doi:10.1162/neco.1997.9.8.1735.

[18] J. Pavlopoulos, V. Kougia, I. Androutsopoulos, A Survey on Biomedical Image Captioning (2019) 26–36, doi:10.18653/v1/w19-1803.

[19] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A.W.M. van der Laak, B. van Ginneken, C.I. Sánchez, A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88, doi:10.1016/j.media.2017.07.005.

[20] M. Tariq, S. Iqbal, H. Ayesha, I. Abbas, K.T. Ahmad, M. Farooq, K. Niazi, Medical Image based Breast Cancer Diagnosis : state of the Art and Future Directions, Expert Syst. Appl. (2020) 114095, doi:10.1016/j.eswa.2020.114095.

[21] X. Liu, F. Hou, H. Qin, A. Hao, Multi-view multi-scale CNNs for lung nodule type classification from CT images, Pattern Recognit 77 (2018) 262–275, doi:10.1016/j.patcog.2017.12.022.

[22] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M.P. Lungren, A.Y. Ng, CheXNet: radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning, (2017) 3–9. http://arxiv.org/abs/1711.05225.

[23] P. Rajpurkar, J. Irvin, R.L. Ball, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C.P. Langlotz, B.N. Patel, K.W. Yeom, K. Shpanskaya, F.G. Blankenberg, J. Seekins, T.J. Amrhein, D.A. Mong, S.S. Halabi, E.J. Zucker, A.Y. Ng, M.P. Lungren, Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists, PLoS Med 15 (2018) 1–17, doi:10.1371/journal.pmed.1002686.

[24] X. Shin, H. Su, F. Xing, Y. Liang, G. Qu, Interleaved Text/Image Deep Mining on a Large-Scale Radiology Database for Automated Image Interpretation, J. Mach. Learn. Res. 17 (2016) 1–31 http://www.jmlr.org/papers/volume17/15-176/15-176.pdf.

[25] W. Gale, L. Oakden-Rayner, G. Carneiro, A.P. Bradley, L.J. Palmer, Producing radiologist-quality reports for interpretable artificial intelligence, (2018) 1–7. http://arxiv.org/abs/1806.00340.

[26] H. Li, J. Weng, Y. Shi, W. Gu, Y. Mao, Y. Wang, W. Liu, J. Zhang, An improved deep learning approach for detection of thyroid papillary cancer in ultrasound images, Sci. Rep. 8 (2018) 1–12, doi:10.1038/s41598-018-25005-7.

[27] B. Gecer, S. Aksoy, E. Mercan, L.G. Shapiro, D.L. Weaver, J.G. Elmore, Detection and classification of cancer in whole slide breast histopathology images using deep convolutional networks, Pattern Recognit 84 (2018) 345–356, doi:10.1016/j.patcog.2018.07.022.

[28] B. Chen, L. Wang, X. Wang, J. Sun, Y. Huang, D. Feng, Z. Xu, Abnormality detection in retinal image by individualized background learning, Pattern Recognit 102 (2020) 107209, doi:10.1016/j.patcog.2020.107209.

[29] H. Xie, D. Yang, N. Sun, Z. Chen, Y. Zhang, Automated pulmonary nodule detection in CT images using deep convolutional neural networks, Pattern Recognit 85 (2019) 109–119, doi:10.1016/j.patcog.2018.07.031.

[30] Y. Cai, Y. Li, C. Qiu, J. Ma, X. Gao, Medical image retrieval based on convolutional neural network and supervised hashing, IEEE Access 7 (2019) 51877–51885, doi:10.1109/ACCESS.2019.2911630.

[31] A. Qayyum, S.M. Anwar, M. Awais, M. Majid, Medical image retrieval using deep convolutional neural network, Neurocomputing. 266 (2017) 8–20, doi:10.1016/j.neucom.2017.05.025.

[32] L. Tsochatzidis, K. Zagoris, N. Arikidis, A. Karahaliou, L. Costaridou, I. Pratikakis, Computer-aided diagnosis of mammographic masses based on a supervised content-based image retrieval approach, Pattern Recognit 71 (2017) 106–117, doi:10.1016/j.patcog.2017.05.023.

[33] A. Ben Abacha, A.G. Seco De Herrera, S. Gayen, D. Demner-Fushman, S. Antani, NLM at ImageCLEF 2017 caption task, in: CEUR Workshop Proc., 2017, p. 1866.

[34] Y. Zhang, X. Wang, Z. Guo, J. Li, ImageSem at ImageCLEF 2018 caption task: image retrieval and transfer learning, in: CEUR Workshop Proc., 2018, p. 2125.

[35] S.S. Azam, M. Raju, V. Pagidimarri, V. Kasivajjala, Q-Map: clinical Concept Mining from Clinical Documents, 560076 (2018). http://arxiv.org/abs/1804.11149.

[36] L. Soldaini, N. Goharian, QuickUMLS: a fast, unsupervised approach for medical concept extraction, Med. Inf. Retr. Work. SIGIR 2016 (2016).

[37] G. Liu, T.-M.H. Hsu, M. McDermott, W. Boag, W.-.H. Weng, P. Szolovits, M. Ghassemi, Clinically Accurate Chest X-Ray Report Generation, (2019). http://arxiv.org/abs/1904.02633.

[38] X. Wang, Y. Peng, L. Lu, Z. Lu, R.M. Summers, TieNet: text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-Rays, Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (2018) 9049–9058, doi:10.1109/CVPR.2018.00943.

[39] L. Wu, C. Wan, Y. Wu, J. Liu, Generative caption for diabetic retinopathy images, in: 2017 Int. Conf. Secur. Pattern Anal. Cybern. SPAC 2017. 2018-Janua, 2018, pp. 515–519, doi:10.1109/SPAC.2017.8304332.

[40] X. Huang, F. Yan, W. Xu, M. Li, Multi-Attention and Incorporating Background Information Model for Chest X-Ray Image Report Generation, IEEE Access 7 (2019) 154808–154817, doi:10.1109/ACCESS.2019.2947134.

[41] S. Liang, X. Li, Y. Zhu, X. Li, S. Jiang, ISIA at the ImageCLEF 2017 image caption task, in: CEUR Workshop Proc., 2017, p. 1866.
[42] S.A. Hasan, Y. Ling, J. Liu, R. Sreenivasan, S. Anand, T.R. Arora, V. Datla, K. Lee, A. Qadir, C. Swisher, O. Farri, PRNA at ImageCLEF 2017 caption prediction and concept detection tasks, in: CEUR Workshop Proc., 2017, p. 1866.
[43] Y. Xue, T. Xu, L.R. Long, Z. Xue, S. Antani, G.R. Thoma, X. Huang, Multimodal recurrent model with attention for automated radiology report generation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Cham, Springer, 2018, pp. 457–466, doi:10.1007/978-3-030-00928-1_52.
[44] Y. Su, F. Liu, M.P. Rosen, UMass at ImageCLEF caption prediction 2018 task, in: CEUR Workshop Proc., 2018, p. 2125.
[45] J. Yuan, H. Liao, R. Luo, J. Luo, Automatic Radiology Report Generation Based on Multi-view Image Fusion and Medical Concept Enrichment, Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 11769 LNCS (2019) 721–729, doi:10.1007/978-3-030-32226-7_80.
[46] O. Pelka, C.M. Friedrich, Keyword generation for biomedical image retrieval with recurrent neural networks, CEUR Workshop Proc (2017) 1866.
[47] D. Lyndon, A. Kumar, J. Kim, Neural captioning for the ImageCLEF 2017 medical image challenges, CEUR Workshop Proc (2017) 1866.
[48] J. Xu, W. Liu, C. Liu, Y. Wang, Y. Chi, X. Xie, X. Hua, Concept detection based on multi-label classification and image captioning approach - DAMO at ImageCLEF 2019, CEUR Workshop Proc 2380 (2019) 9–12.
[49] X. Li, R. Cao, D. Zhu, Vispi: automatic Visual Perception and Interpretation of Chest X-rays, (2019). http://arxiv.org/abs/1906.05190.
[50] C.Y. Li, X. Liang, Z. Hu, E.P. Xing, Knowledge-Driven Encode, Retrieve, Paraphrase for Medical Image Report Generation, in: Proc. AAAI Conf. Artif. Intell., 33, 2019, pp. 6666–6673, doi:10.1609/aaai.v33i01.33016666.
[51] Y. Zhang, D.Y. Ding, T. Qian, C.D. Manning, C.P. Langlotz, Learning to Summarize Radiology Findings, (2009).
[52] G. Spinks, M.F. Moens, Generating text from images in a smooth representation space, CEUR Workshop Proc (2018) 2125.
[53] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 8689 LNCS (2014) 818–833, doi:10.1007/978-3-319-10590-1_53.
[54] K. Papineni, S. Roukos, T. Ward, W. Zhu, Y. Heights, I.B.M. Research Report Bleu : a Method for Automatic Evaluation of Machine Translation, Science (80-.) 22176 (2001) 1–10, doi:10.3115/1073083.1073135.
[55] A. Lavie, A. Agarwal, METEOR: an automatic metric for MT evaluation with high levels of correlation with human judgments, Proc. Second Work. Stat. Mach. Transl. 0 (2007) 228–231 http://acl.ldc.upenn.edu/W/W05/W05-09.pdf#page=75.
[56] C.Y. Lin, Rouge: a package for automatic evaluation of summaries, Proc. Work. Text Summ. Branches out (WAS 2004) (2004) 25–26 papers2://publication/uuid/5DDA0BB8-E59F-44C1-88E6-2AD316DAEF85.
[57] R. Vedantam, C.L. Zitnick, D. Parikh, CIDEr: consensus-based image description evaluation, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 07-12-June, 2015, pp. 4566–4575, doi:10.1109/CVPR.2015.7299087.
[58] P. Anderson, B. Fernando, M. Johnson, S. Gould, SPICE: semantic propositional image caption evaluation, Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 9909 LNCS (2016) 382–398, doi:10.1007/978-3-319-46454-1_24.
[59] M. Kilickaya, A. Erdem, N. Ikizler-Cinbis, E. Erdem, *Re*-evaluating automatic metrics for image captioning, 15th Conf. Eur. Chapter Assoc. Comput. Linguist. EACL 2017 - Proc. Conf. 1 (2017) 199–209. https://doi.org/10.18653/v1/e17-1019.
[60] D. Elliott, F. Keller, Comparing automatic evaluation measures for image description, in: 52nd Annu. Meet. Assoc. Comput. Linguist. ACL 2014 - Proc. Conf., 2, 2014, pp. 452–457, doi:10.3115/v1/p14-2074.
[61] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: a neural image caption generator, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 07-12-June, 2015, pp. 3156–3164, doi:10.1109/CVPR.2015.7298935.
[62] K. Xu, J.L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R.S. Zemel, Y. Bengio, Show, attend and tell: neural image caption generation with visual attention, in: 32nd Int. Conf. Mach. Learn. ICML 2015, 2015, pp. 2048–2057.
[63] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, Commun. ACM. 60 (2017) 84–90, doi:10.1145/3065386.
[64] Y. Lecun, L. Bottou, Y. Bengio, P. Ha, LeNet, Proc. IEEE. (1998) 1–46. https://doi.org/10.1109/5.726791.
[65] R. Kiros, R. Salakhutdinov, R.S. Zemel, Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models, (2014) 1–13. http://arxiv.org/abs/1411.2539.
[66] M. Lin, Q. Chen, S. Yan, Network in network, in: 2nd Int. Conf. Learn. Represent. ICLR 2014 - Conf. Track Proc., 2014, pp. 1–10.
[67] S. Ren, K. He, R. Girshick, J. Sun, R.-.C.N.N. Faster, Towards Real-Time Object Detection with Region Proposal Networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2017) 1137–1149, doi:10.1109/TPAMI.2016.2577031.
[68] Y. Kim, K. Zhang, A.M. Rush, Y. Lecun, Adversarially Regularized Autoencoders, (2018).
[69] I.J. Goodfellow, J. Pouget-abadie, M. Mirza, B. Xu, D. Warde-farley, Generative Adversarial Nets (2014) 1–9 n.d.
[70] S. Ioffe, C. Szegedy, Batch Normalization : accelerating Deep Network Training by Reducing Internal Covariate Shift, (2015) (n.d.).
[71] A. Graves, Framewise phoneme classification with bidirectional LSTM and other neural network architectures, 18 (2005) 602–610. https://doi.org/10.1016/j.neunet.2005.06.042.

Hareem Ayesha has completed her undergraduate from Department of Computer Science, Bahauddin Zakariya University, Multan in 2018. At present she is doing her Master of Science in Computer Science from Department of Computer Science, Bahauddin Zakariya University, Multan. Her research interests include Artificial Intelligence, Computer Vision and Medical Image Analysis.

Sajid Iqbal has completed his BSCS from Department of Computer Science, Bahauddin Zakariya University, Multan, Pakistan, in 2002, and Master of Science in Computer Science from National University of Computer and Emerging Sciences, Lahore, Pakistan, in 2003. After this he has been associated with different institutes of higher education and taught courses at undergraduate and graduate level. He completed his Ph.D from University of Engineering and Technology, Lahore, Pakistan. Currently he is working as an Assistant Professor at Department of Computer Science, Bahauddin Zakariya University, Multan, Pakistan. He is published more than 15 research papers in well reputed international journals. His-research areas include deep learning, medical image analysis and natural language processing.

Mehreen Tariq has completed her undergraduate from Department of Computer Science, Bahauddin Zakariya University, Multan in 2018. Currently she is enrolled in Master of Science in Computer Science program as a research student. She is interested in Artificial Intelligence as her research domain. Sub fields of her research interest include medical image processing and computer vision.

Muhammad Abrar has completed his BSIT from Department of Computer Science University of Education, Multan, Pakistan. Currently he is enrolled in his Master of Science in Computer Science degree program at Department of Computer Science, Nawaz Sharif Agriculture University, Multan, Pakistan. His-research interests include image processing, machine learning and Artificial Intelligence in general.

Muhammad Sanaullah has been working as an Assistant Professor in the Department of Computer Science, Bahauddin Zakariya University, Multan, Pakistan. His-current research focuses on the use of Semantic Web Technologies in the field of Machine Learning, Data Mining and IoTs.

Ishaq Abbas has completed his undergraduate from Bahauddin Zakariya University, Multan in 2018. At present he is pursuing his research based graduate program (Master of Science in Computer Science) at Department of Computer Science, Bahauddin Zakariya University, Multan. Areas of his research interest are deep learning and computer vision.

Amjad Rehman received the Ph.D. degree in image processing and pattern recognition from Universiti Teknologi Malaysia, Malaysia, in 2010. During his Ph.D., he proposed novel techniques for pattern recognition based on novel features mining strategies. He is currently conducting research under his supervision for three Ph.D. students. He is the author of dozens of papers published in international journals and conferences of high repute. His-keen research includes information security, data mining and documents analysis, recognition.

Muhammad Farooq Khan Niazi Dr. Niazi completed his MBBS in 1985 and has served in multiple health care units and institutes. He completed his FCPS fellowship in 1999. He has more than 24 years practice experience in the field of diagnostic radiology. He has also served as senior faculty member at College of Physicians and Surgeons Pakistan. Presently, he is serving as a professor of radiology/ head of department at Bakhtawar Amin Memorial Trust Hospital Multan, Pakistan.