BioMedical Engineering OnLine

REVIEW Open Access

A survey on automatic generation of medical imaging reports based on deep learning

Ting Pang[1*], Peigao Li[1] and Lijie Zhao[1]

*Correspondence:
pt@xxmu.edu.cn

[1] Center of Network and Information, Xinxiang Medical University, Xinxiang 453000, China

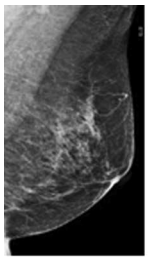
Abstract

Recent advances in deep learning have shown great potential for the automatic generation of medical imaging reports. Deep learning techniques, inspired by image captioning, have made significant progress in the field of diagnostic report generation. This paper provides a comprehensive overview of recent research efforts in deep learning-based medical imaging report generation and proposes future directions in this field. First, we summarize and analyze the data set, architecture, application, and evaluation of deep learning-based medical imaging report generation. Specially, we survey the deep learning architectures used in diagnostic report generation, including hierarchical RNN-based frameworks, attention-based frameworks, and reinforcement learning-based frameworks. In addition, we identify potential challenges and suggest future research directions to support clinical applications and decision-making using medical imaging report generation systems.

Keywords: Medical imaging reports, Automatic generation, Image captioning, Deep learning

Introduction

As we all know, a detailed explanation of medical images such as CT (computed tomography), ultrasound, MRI (magnetic resonance imaging), or pathological imaging must be conducted by professional physicians or pathologists who write a diagnostic report for each patient. An example of such a report can be seen in Fig. 1. Although one report may seem simple, containing only indications, findings and impression, there are many patients with unforeseen abnormal medical images. Therefore, analyzing and depicting textual reports, which require skilled experience, can be a time-consuming and stressful task for professionals. Automatic diagnostic report generation from medical images is an indispensable trend to reduce this workload. In addition, while deep learning, with its advantage of end-to-end processing, has emerged on a large scale in recent medical diagnosis studies, the non-interpretable network and non-standardized evaluation make deep learning like a black box. Teaching machines to automatically write diagnostic reports is a semantic and effective way to support the interpretability of deep learning models [1]. Hence, it is essential to explore the automatic diagnosis of images and the generation of reports to improve the interpretability of deep learning.



Indication: High risk inherited family.

Findings: Breast density category: BIRADS c. No masses, significant clustered or pleomorphic microcalcifications and new areas of architectural distortion.

Impression: Normal mammogram, BIRADS 1.

Fig. 1 One simple example of mammography report

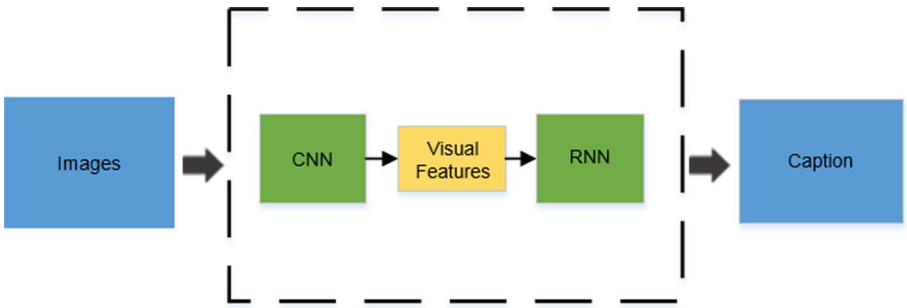


Fig. 2 Illustration of the CNN–RNN-based image captioning framework

The automatic generation of diagnostic reports is inspired by image captioning [2], which combines computer vison (CV) and natural language processing (NLP) to provide a comprehensive understanding of medical images. Traditionally, image captioning was achieved through report retrieval [3] and template-based generation [4]. However, these conventional methods are limited in their ability to produce flexible and comprehensive textual descriptions that can be applied to new images. Recent progress in deep learning has led to significant advancements in image captioning. In this research, we focus on medical report generation based on deep learning. Essentially, the paradigm follows a typical encoder–decoder architecture [5–7]. It leverages visual features obtained from Convolution Neural Network (CNN, encoder) to generate descriptions of given images through Recurrent Neural Network (RNN, decoder) [8], as shown in Fig. 2.

However, generating diagnostic reports is a challenging task due to the complexity and diversity of objects in medical images. In practice, the values obtained via the activation function at one suitable layer of the objects recognition CNN are considered as the visual feature vector [9]. Moreover, variations of RNN, such as long–short-term memory (LSTM) [10] and gated recurrent unit (GRU) [11], that contain different controlling gates capable of learning information from a long time ago, are frequently employed in effectively capturing the semantics of image captioning tasks. In addition, more recent works focus on generating long-form text instead of single sentences [12, 13]. Attention mechanisms that focus on salient parts have been widely used in image captioning to provide visual explanations for the rationale of deep learning networks [14–17]. Reinforcement Learning [18] (RL) and Generative Adversarial Networks [19] (GAN) have also been widely proposed in image captioning [20] due to their recent success.

To date, some scholars have explored the automatic generation of medical reports using image captioning methods move forward, see the basic framework in Fig. 3. The first application of deep learning in medical imaging report generation was conducted by Shin et al. [21] in 2016. They developed a CNN–RNN network that effectively predicted only annotated tags (e.g., location, severity and affected organs) from chest X-ray images. They tested both LSTM and GRU and improved the results by considering joint image/text contexts using account using a recurrent neural cascade model. LSTM has been more widely used and studied in the literature, and has achieved state-of-the-art results in many tasks. However, GRU is gaining popularity due to its simpler architecture and faster training time compared to LSTM. Subsequently, in further research on medical image captioning, LSTM will be used as the core framework of RNN.

The primary aim of this manuscript is to present a systematic review of studies on deep learning-based medical imaging reports generation. The survey provides readers with a comprehensive understanding of the field of deep learning in automatic diagnostic reports generation, and to offer clinical treatment management suggestions for medical imaging reports generation exploiting deep learning. The survey also lays the foundation for innovation to increase the richness of this field. To summarize, this work contributes in three ways: (1) it focuses on the clinical value of deep learning-based diagnostic reports generation, providing suggestions for clinical decision making and reducing the workload of radiologists; (2) it organizes and explains the current works in detail, proving that automatic writing diagnostic reports can improve the interpretability of deep learning in medical imaging area; and (3) it provides comprehensive references and identifies new trends for researchers in this field. This paper is the first overview of medical report generation based on deep learning, with a focus on improving interpretability of deep learning and its clinical value.

This paper is structured as follows: in "Overview and analysis" section, we provide a comprehensive summary and analysis of the current state of deep learning applied in medical imaging reports generation, covering aspects, such as data sets, architectures, applications and evaluations based on the retrieved studies. In "Discussion and future"

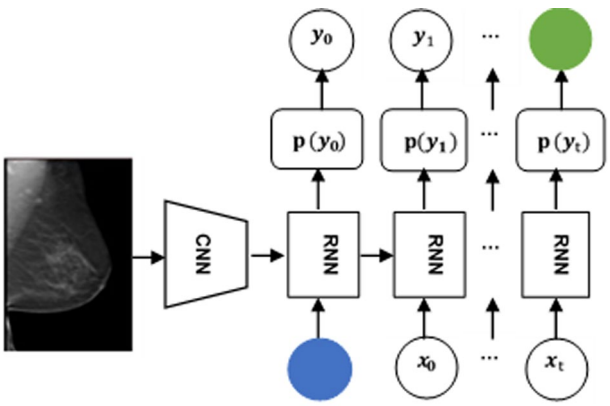


Fig. 3 Illustration of the CNN–RNN-based framework for diagnostic report generation. The variable "*t*" represents time, "*x*" denotes the input layer, "*y*" represents the output layer, and "*p*(*y*)" denotes the probability of output

section, we discuss potential challenges and future directions to serve as a reference for further studies in this field. Finally, in "Conclusion" section, we provide brief conclusions.

Overview and analysis

The encoder–decoder framework, which combines image-text embedding models with multimodal neural language models, was first introduced by [22]. The framework encodes visual data, projecting it into the embedded space composed of RNN hidden states that encode text data by optimizing the pairwise sorting loss. In the embedding space, a structure-content neural language model is used to decode the visual features, based on the feature vectors of context words, to form sentences. An example of the whole framework can be seen in Fig. 4.

Within the framework described above, image captioning is defined as the probability of generating a sentence based on an input image (Eq. 1):

$$S^* = {}_S{}^{argmax} \prod P(S_t|I, S_0, ..., S_{t-1}; \theta) \tag{1}$$

where I is the input image, θ is the model parameter. A sentence S equals to a sequence of words $S_0, ..., S_{t-1}$.

Vinyals et al. use a LSTM neural network [8] to model $P(S_t|I, S_0, ..., S_{t-1}; \theta)$ as hidden state h_t, which can be updated as (Eq. 2)

$$h_{t+1} = f(h_t, x_t) \tag{2}$$

where x_t is the input to the LSTM neural network. In the first unit, x_t is an image feature, while in other units x_t is a feature of previously predicated context words. The model parameter θ is obtained by maximizing the likelihood of sentence-image pairs in the training set. Through the training model, the possible output word sequences can be predicted by sampling or beam search.

To generate descriptions closely related to image contents, Jia et al. (2016) extracted semantic information from images and added it to each unit of the LSTM in the process

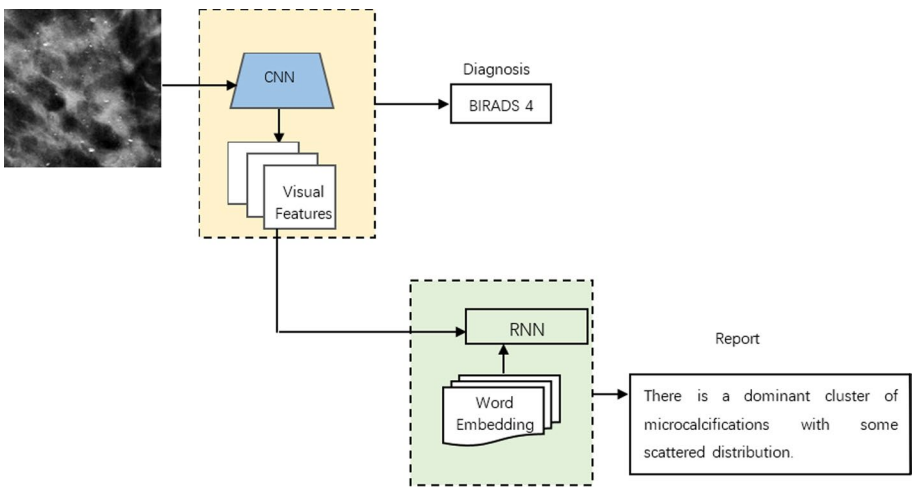


Fig. 4 Medical report generation example of the encoder–decoder framework

of sentence generation [23]. The original forms of the memory unit and gate of an LSTM unit [24] are defined as (Eqs. 3, 4, 5, 6, 7)

$$i_l = \sigma(W_{ix}x_l + W_{im}m_{l-1}) \tag{3}$$

$$f_l = \sigma(W_{fx}x_l + W_{fm}m_{l-1}) \tag{4}$$

$$o_l = \sigma(W_{ox}x_l + W_{om}m_{l-1}) \tag{5}$$

$$c_l = f_l \odot c_{l-1} + i_l \odot h(W_{cx}x_l + W_{cm}m_{l-1}) \tag{6}$$

$$m_l = o_l \odot c_l \tag{7}$$

where variables i_l, f_l and o_l, respectively, denotes input gate, forget gate, output gate of a LSTM cell, c_l and m_l denotes the state and hidden state of the memory cell unit, $\sigma(\cdot)$ and $h(\cdot)$ are non-linear functions, x_l is the input, W are model parameters, and $\odot$ stands for an elementwise multiplication operation.

Aiming to utilize high-level semantic information for image captioning, Qi et al. (2016) incorporate a set of semantic attributes from the training sentences which are seen as visual concepts into the encoder–decoder framework [25]. In the region-based multi-label classification framework [26], a CNN-based multi-classifier is trained for each attribute. By training the semantic attribute classifiers, the image I can be encoded as a prediction vector $V_{att}(I)$ giving the probability of each attribute appearing in the image. Then, a LSTM is deployed as decoder to generate a sentence describing the contents of the image based on the representation. In this case, the image captioning problem can be rephrased as (Eq. 8)

$$S^* = {}_S{}^{argmax} P(S|V_{att}(I); \theta) \tag{8}$$

where I is the input image, θ is the model parameter, S is a sentence.

Data sets

The automatic generation of medical imaging reports based on deep learning requires a large data set for training. In this section, we introduce frequently used public data sets and some typical private data sets.

The current public data sets have greatly contributed to the development of deep learning for medical imaging report generation. The most commonly used databases consist of images and reports from the United States and Europe, with chest radiographs being the predominant data set. Some examples of these data sets include Indiana University Chest XRay (IU X-Ray) [27], ChestX-ray14 [28], CheXpert [29], MIMIC Chest X-ray (MIMIC-CXR) [30], CX-CHR [31], PadChest [32], as shown in Table 1.

The IU X-Ray is a set of chest X-ray images paired with their corresponding diagnostic reports. The data set contains 7470 images (6470:500:500) and 3955 report. Each report consists of the following sections: impression, findings, tags, comparison, and indication. On average, each image is associated with 2.2 tags, 5.7 sentences, and each sentence contains 6.5 words. About 70% of the automatic report generation work are from

Table 1 Common data set of medical imaging report generation

Data set	Description	Image	Report	Link
IU X-Ray	Chest X-ray images of lung diseases	7470	3955	http://openi.nlm.nih.gov/
ChestX-ray14	14 kinds of lung diseases	112,120	–	https://nihcc.app.box.com/v/ChestXray-NIHCC
CheXpert	Chest radiographs of 65,240 patients with lung diseases	224,316	–	https://stanfordmlgroup.github.io/competitions/chexpert/
MIMIC-CXR	227,835 radiographic studies in DICOM format	377,110	227,835	https://physionet.org/content/mimic-cxr/2.0.0/
CX-CHR	Chest X-ray images with Chinese reports of 35,609 patients	45,598	–	–
PadChest	Chest X-ray data set obtained from 67,000 patients	160,000	109,931	https://bimcv.cipf.es/bimcv-projects/padchest/
PEIR Gross	Radiology teaching images	4,000	4000	https://peir.path.uab.edu/library/index.php?/category/106
DDSM	Normal, benign, and malignant mammography studies	2620	–	http://marathon.csee.usf.edu/Mammography/Database.html

these public data sets, where IU X-Ray takes up the biggest fraction due to its large numbers and comprehensive annotation.

ChestX-ray14 is provided by the national institute of health (NIH). It comprises 112,120 frontal-view X-ray images of 30,805 (collected from the year of 1992 to 2015) unique patients with the common disease labels, mined from the text radiological reports. The database contains 14 kinds of lung diseases (atelectasis, consolidation, infiltration, pneumothorax, edema, emphysema, fibrosis, effusion, pneumonia, pleural thickening, cardiac hypertrophy, nodules, swelling and hernia).

The CheXpert data set contains 224,316 chest radiographs of 65,240 patients with both frontal and lateral views available. The task is to do automated chest X-ray interpretation, which features uncertainty labels and radiologist-labeled reference standard evaluation sets.

MIMIC-CXR is a large publicly available data set of chest radiographs in DICOM format with free-text radiology reports. The data set contains 377,110 images corresponding to 227,835 radiographic studies performed at the Beth Israel Deaconess Medical Center in Boston. The data set is intended to support a wide body of research in medicine including image understanding, natural language processing, and decision support.

CX-CHR is a proprietary internal data set of chest X-ray images with Chinese reports collected from a professional medical institution for health checking. The data set consists of 35,609 patients and 45,598 images. Each patient has one or multiple chest X-ray images in different views, such as poster anterior and lateral, and a corresponding Chinese report.

PadChest is a labeled large-scale, high resolution chest X-ray data set for the automated exploration of medical images along with their associated reports. This data set includes more than 160,000 images obtained from 67,000 patients that were interpreted and reported by radiologists at Hospital San Juan Hospital (Spain) from 2009 to 2017, covering six different position views and additional information on image acquisition and patient demography. The reports were labeled with 174 different radiographic findings, 19 differential diagnoses and 104 anatomic locations organized as a hierarchical

taxonomy and mapped onto standard Unified Medical Language System (UMLS) terminology.

Apart from the chest radiographs, there are some other medical images. Such as PEIR Gross, Digital Database for Screening Mammography (DDSM) [33], etc. PEIR Gross is a collection of over 4,000 curated radiology teaching images, which are created by the University of Alabama for medical education. It contains sentence-level descriptions of 20 different body parts, including the abdomen, adrenal, aorta, breast, chest, head, kidneys, etc. DDSM contains 2620 scanned films of normal, benign, and malignant mammography studies with verified pathology information. It is supported by the University of South Florida and it has been widely used by researchers due to its scale and ground truth validation. Moreover, researchers have trained their deep learning frameworks on several privately owned data sets.

However, private medical imaging data sets are less common. Collecting private medical images can be difficult due to patient confidentiality and data privacy concerns, as well as the laborious effort required for properly indexing, storing, and annotating the images. In addition, image attributes such as cropped image size, format, data source, and number of samples for training and testing can greatly impact the final results [27] [28].

Methods

Hierarchical RNN-based framework

As illustrated in Fig. 5, a medical imaging report typically consists of at least one paragraph consisting of several sentences, which can be much longer for abnormal diseases. To address this challenge, Jing et al. proposed a hierarchical LSTM consisting of a sentence LSTM and a word LSTM for generating long chest X-ray reports, inspired by the hierarchical RNN for image captioning proposed by Krause et al. [12]. The single-layer sentence LSTM determines the number of sentences for medical reports using visual features as inputs and generates the topic vector for each sentence, which is then passed to the two-layer word LSTM. The word LSTM generates fine-grained words and descriptions based on the topics for each sentence, which are concatenated to form the final medical report paragraph (see the hierarchical LSTM report generation model in Fig. 5). Harzig et al. also employed hierarchical LSTM to produce diagnostic reports for chest X-ray, and to address data bias, they innovatively proposed dual word LSTMs, including an abnormal word LSTM and a normal word LSTM, which are trained when the label is abnormal and normal [35]. They also set an abnormal sentence predictor to determine whether to use the sentences generated by the dual word LSTM. To address the limited

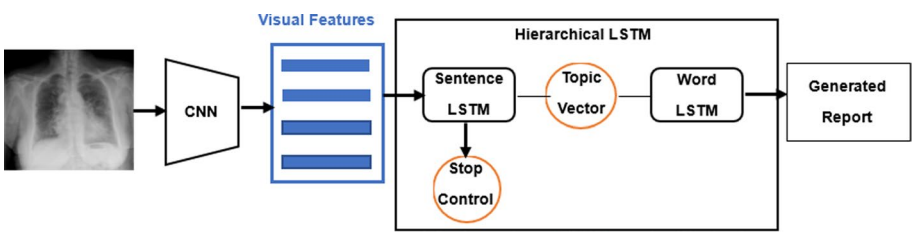


Fig. 5 Hierarchical RNN-based framework for medical report generation

availability of pairs of medical images and reports, Yuan et al. synthesized visual features by taking advantage of multi-view chest X-ray images at the sentence-level LSTM to ensure cross-view consistency [36]. Furthermore, medical concepts based on reports were extracted and merged with respective decoding steps by the word-level LSTM.

Attention-based framework

Recently, attention-based medical image captioning frameworks have been used to provide meaningful embeddings and improve the interpretability of deep learning processes for report generation (Fig. 6). Zhang et al. built a MDNet for bladder cancer diagnosis that combines an image model and a language model, using an improved attention mechanism to enhance image alignment and generate sharper joint image/report attention maps [37]. Wang et al. proposed TieNet, a multi-level attention mechanism that fuses visual attention and text-based attention into a CNN–RNN model to highlight important report and image representations of chest X-ray patients [38]. Lee et al. designed a justification generator to explain the diagnostic decision of breast masses, utilizing attention to obtain visual pointing maps and an LSTM to generate diagnostic sentences [39]. Li et al. adopted an attentive LSTM that takes either the original chest X-ray image or the cropped abnormal ROI as input and generates the entire report [40].

Reinforcement learning-based framework

Motivated by the successful application of reinforcement learning in deep learning, some researchers have attempted to employ RL for optimizing medical imaging report generation, as shown in the basic framework in Fig. 7. RL is formed by agents that learn an optimal policy for better decision-making by receiving rewards from the environment at a given state. Jing et al. proposed a novel Cooperative Multi-Agent System (CMAS) consisting of Planner (PL), Abnormality Writer (AW), and Normality Writer (NW) with one reward module to capture the bias between normality and abnormality for generating more accurate chest X-ray reports [41]. PL determines whether the area has lesions, and AW or NW generates a sentence based on the result given by PL. Similarly, Liu et al. used a final fine-tuned RL containing natural language generation reward and clinically coherent reward to optimize a hierarchical CNN–RNN-based model for clinical accuracy and readability of chest X-ray reports [42].

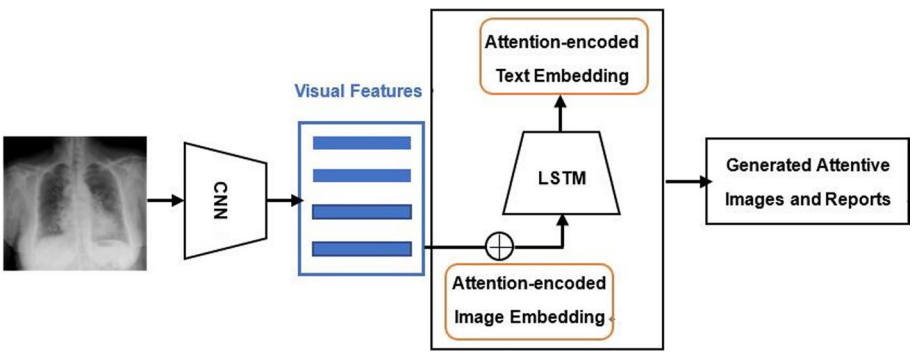


Fig. 6 Attention-based framework for medical report generation

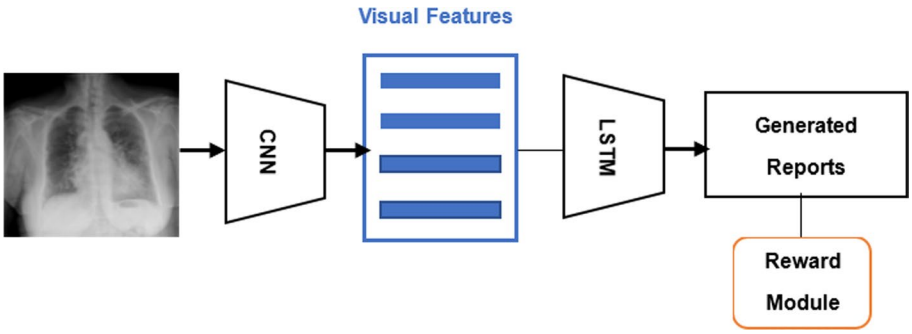


Fig. 7 Reinforcement learning-based framework for medical report generation

Other-related works

Labeling pairs of medical images and reports is a tedious task for professionals. To address this issue, Han et al. proposed a weakly supervised framework that combines symbolic program synthesis theory and deep learning. This framework uses object-level annotations, without requiring radiologist-level report annotations, to generate unified reports [43]. Similarly, Xue et al. developed a recurrent image captioning model that generates the findings of a medical report sentence by sentence, where each successive sentence is based on multimodal inputs, including the original images and the previous sentence [44]. Zeng et al. introduced a coarse-to-fine ultrasound image captioning ensemble model that helps doctors automatically generate high-quality annotated ultrasound reports [45].

Applications

The application of automatic generation of medical imaging reports has a wide range of potential benefits beyond assisting diagnosis and lightening workload. For instance, generating accurate and comprehensive reports can improve patient care by providing more informed treatment decisions. In addition, the vast amounts of data generated by medical imaging can be utilized for medical research and advancements in the field. However, efficient and accurate annotation and labeling is required, which can be facilitated by automatic report generation. In summary, the use of deep learning for automatic generation of medical imaging reports has significant potential to greatly benefit the healthcare industry.

Assisting diagnosis

Some studies have employed a combination of language models (such as LSTM) and image models (such as CNN) to improve the accuracy of diagnostic conclusions. These models leverage the semantic knowledge of medical images obtained from diagnostic reports to provide an interpretable prediction mechanism. To ensure the reliability of the machine learning system's decisions, it is important to open the black box of deep learning and increase understanding of the reasoning behind the decisions [46]. All the studies reviewed attempt to present semantically and visually interpretable results during the diagnosis process [46–49].

Lighten workload

In addition to these modalities and categories of diseases, automatic generation of medical imaging reports has also been explored in other areas such as MRI, CT scans, and PET scans for various diseases such as lung cancer, brain tumors, and cardiovascular diseases. The tedious process of preparing reports can be a significant burden on radiologists and can lead to errors or delays in patient care. By automating this process, radiologists can focus on more complex tasks and improve patient outcomes. Furthermore, the generated reports can provide valuable insights for medical research and contribute to the development of new treatment options.

Evaluations

BLEU [50], ROUGE [51], METEOR [52] and CIDER [53] are commonly used evaluation metrics for medical image report generation, which are adapted from machine translation and text summarization.

BLEU (Bilingual Evaluation Understudy) measures the similarity between the generated report and the ground truth report by calculating the overlap of word n-grams. BLEU-1 measures the overlap of unigrams (i.e., single words), while BLEU-2, -3, and -4 consider bigrams, trigrams, and quadrigrams, respectively. To account for short generated reports, a penalty is added to the score. BLEU is easy to calculate and interpret, and it has been shown to correlate well with human judgments of text quality. However, BLEU only considers surface-level similarities between the generated and reference texts, and it does not take into account the semantic content or coherence of the generated text.

METEOR (Metric for Evaluation of Translation with Explicit ORdering) extends BLEU-1 by adopting F-score of precision and recall, with a bias towards recall, and utilizing Porter stemmer and WordNet. To account for longer subsequences, it includes a penalty of up to 50% when there are no common n-grams between machine-generated descriptions and references. METEOR takes into account both surface-level and semantic similarities between the generated and reference texts. It also has a built-in mechanism for handling synonyms and paraphrases. Like BLEU, METEOR does not account for the coherence or overall quality of the generated text.

ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation—Longest Common Subsequence) measures the longest common subsequence between the machine-generated description and the reference human description, and calculates its ratio to the reference size (ROUGE-L recall), generated description (ROUGE-L precision), or a combination of the two (ROUGE-L F-measure). ROUGE-L takes into account the semantic content and coherence of the generated text, and it has been shown to correlate well with human judgments of text quality. However, ROUGE-L only considers a single metric, the longest common subsequence, and it may not capture all aspects of text quality.

CIDER (Consensus-based Image Description Evaluation) measures the cosine similarity between n-gram TF–IDF (Term Frequency–Inverse Document Frequency) representations of the generated report and the reference report (words are also stemmed). The calculation is done from single gram to 4 g and the average is returned as the final evaluation score. The rationale behind using TF–IDF is to reward frequent words and penalize common words (such as stop words). CIDER takes into account both surface-level

and semantic similarities between the generated and reference texts. It also has been shown to correlate well with human judgments of text quality for image captioning tasks. However, CIDER may not be suitable for tasks other than image captioning, and it is computationally more expensive than other evaluation metrics.

Automatic generation of medical reports using deep learning is still an emerging area with many challenges. We conducted a search of 31 relevant papers and compiled detailed implementation information in Table 2.

Discussion and future

Despite the significant progress made in medical imaging report generation based on deep learning, this section aims to highlight the unresolved issues and present future research directions in this area for further development.

Balanced data set

Deep learning has shown great potential in big data analytics, but in the field of medical imaging report generation, there are still many challenges to be addressed. One major issue is the imbalanced nature of available data sets. There is a lack of public databases that include a variety of image modalities, such as pathology, ultrasound, and magnetic resonance imaging (MRI). In addition, private data sets are often arbitrary in terms of number, size, and format, which makes it difficult to compare results across studies. Another challenge for both private and public data set is the annotation of images, as clinical radiologists may not always be available due to the labor-intensive and time-consuming nature of the task. The use of imbalanced data sets for training neural networks can lead to biased diagnostic report generation. To address these challenges, we need to establish public databases with a variety of image modalities, as well as develop private data sets to address the limitations of medical images and complex annotations. Private data sets can be useful for clinical practice, such as combining different imaging modalities and diagnostic reports from various sources.

Clinical application

Clinical decision-making is critical in patient management and care, and errors in medical imaging reporting can lead to serious consequences. Therefore, improving the accuracy of medical reports is crucial. While deep learning has shown great potential in this field, there is still a significant research gap in the domain of diagnostic report generation. Many studies focus on improving the final performance, but we should also pay attention to the deep features obtained by deep learning and consider the unique characteristics of different diseases for accurate report generation. By doing so, we can enhance the practical application value of deep learning in clinical decision-making.

Unified evaluation

In many studies, the technical details of experiments were not described in sufficient detail. The selection of measurement indicators and baseline methods was often

Table 2 Studies conducted for medical report generation based on deep learning

References	Data (Images, reports)	Architecture	BLEU-1	BLEU-2	BLEU-3	BLEU-4	MET-EOR	ROU-GE	CIDEr
Shin et al. 2016 [21]	OpenI (7470, 3955)	CNN–RNN	0.972	0.671	0.149	0.028	–	–	–
Zhang et al. 2017 [37]	Bladder Cancer (1000, 5000)	CNN–LSTM–ATT	0.912	0.829	0.75	0.677	0.396	0.701	0.0204
Jing et al.2017 [34]	IU X-Ray (7470, 7470)	CNN–HLSTM–ATT	0.517	0.386	0.306	0.247	0.217	0.447	0.327
Wang et al. 2018 [38]	ChestX-ray14 (—, —)	CNN–LSTM–ATT	0.2860	0.1597	0.1038	0.0736	0.1076	0.2263	–
Xue et al. 2018 [44]	IU X-Ray (7470, 7470)	Recurrent CNN–LSTM–ATT	0.464	0.358	0.270	0.195	0.274	0.366	–
Han et al. 2018 [43]	Lumbar Spinal MRI (253, 253)	Weakly Supervised CNN–LSTM	–	–	–	–	–	–	–
Tian et al. 2018 [54]	CT (—, —)	CNN–LSTM	–	–	–	0.766	–	–	–
Zeng et al. 2018 [45]	Ultrasound Image (—, —)	CNN–LSTM	0.22	0.13	0.09	-	0.10	0.39	0.90
Ma et al. 2018 [55]	Pathology (—, —)	CNN–LSTM	–	–	–	–	–	–	–
Harzig et al. 2019 [35]	IU X-Ray (7470, 3955)	CNN–HLSTM–DualLSTM–ATT	0.373	0.246	0.175	0.126	0.163	0.315	0.359
Yuan et al. 2019 [36]	CheXpert (6248, -)	Muti-view CNN–LSTM–ATT–Medical Concepts	0.529	0.372	0.315	0.255	0.343	0.453	-
Lee et al. 2019 [39]	DDSM FFDM2.0 (605, 605)	CNN–LSTM–ATT	0.4070	0.2296	0.1354	0.0871	-	0.2650	0.1366
Liu et al. 2019 [42]	MIMIC-CXR(327,281, 141,783)	CNN–HLSTM–RL	0.313	0.206	0.146	0.103	0.146	0.306	1.046
Jing et al. 2019 [41]	CX-CHR (—, —)	CMAS–RL	0.428	0.361	0.323	0.290	–	0.504	2.968
Gale et al. 2019 [56]	Frontal Pelvic X-rays (50,363, -)	CNN–LSTM–ATT	0.919	0.838	0.761	0.677	–	–	–
Hasan et al. 2019 [57]	Biomedical Images(164,614, -)	CNN–LSTM	0.3211	–	–	–	–	–	–
Sun et al. 2019 [58]	INbreast (—, —)	CNN–LSTM	–	–	–	–	–	–	–
Xie et al. 2019 [59]	–	CNN–LSTM–ATT	–	–	–	–	–	–	–
Li et al. 2019 [40]	IU X-Ray (7470, 7470)	CNN–LSTM–ATT	0.419	0.280	0.201	0.150	-	0.371	0.553
Yin et al. 2020 [60]	Two image-paragraph pair data sets	Hierarchical RNN	–	–	–	–	–	–	–
Pino et al. 2020 [61]	IU X-Ray (7470, 7470)	CNN–LSTM–ATT	0.361	0.226	0.152	0.106	-	0.314	0.187
Zeng et al. 2020 [62]	Ultrasound image	CNN–LSTM	–	–	–	–	–	–	–
Xu et al. 2020 [63]	IU X-Ray (7470, 7470) and MIMIC-CXR	Reinforce CNN–LSTM	0.412	0.279	0.206	0.157	0.179	0.342	0.411
Singh et al. 2021 [64]	IU X-Ray (—, —)	CNN–LSTM-	23.07	11.86	7.05	4.75	11.11	23.15	19.78
Yang et al. 2021) [65]	Ultrasound image	Adaptive Multimo-dal ATT	–	–	–	–	–	–	–

Table 2 (continued)

References	Data (Images, reports)	Architecture	BLEU-1	BLEU-2	BLEU-3	BLEU-4	MET-EOR	ROU-GE	CIDEr
Najdenkoska et al. 2021 [66]	IU X-Ray (7470, 7470) and MIMIC-CXR	CNN–LSTM–ATT	–	–	–	–	–	–	–
Oa et al. 2021 [67]	IU X-Ray (7470, 7470)	Condition GPT2	0.387	0.245	0.166	0.111	0.164	0.289	0.257
Liu et al. 2021 [68]	COVID-19 cases (1104, 368)	Medical visual language BERT	–	–	–	–	–	–	–
Han et al. 2021 [69]	spinal image data set	Neural-symbolic learning (NSL) framework	–	–	–	–	–	–	–
Wu et al. 2022 [70]	skin pathological image data set (1147, 1147)	CNN–LSTM–ATT	–	–	–	–	–	–	–
Chang et al. 2022 [71]	lung CT scans (458, 458)		–	–	–	–	–	–	–

ATT Attention

arbitrary, resulting in a lack of standardization in the evaluation process. Most researchers focused on metrics, such as BLEU, ROUGE, METEOR, and CIDER, which are commonly used in natural image evaluation but may not be appropriate for medical imaging reports. To improve the evaluation process, it is necessary to design more specific metrics in the medical domain to better evaluate the accuracy and interpretability of the generated reports.

Interdisciplinary background

The progress in deep learning for medical imaging report generation is hindered by the lack of collaboration between experts from different fields. Many medical professionals lack the technical expertise to design and code deep learning models, while engineering and computer science specialists may not have sufficient knowledge of medical imaging and complex clinical applications. Better communication and a closer working relationship between these fields are essential to advance deep learning for clinically useful applications in medical imaging report generation.

Conclusion

Automatic generation of diagnostic reports from medical images can significantly reduce the workload of report writing. In addition, using semantic information to express visual features can improve the interpretability of deep learning-based models. This paper presents a survey of recent studies on deep learning-based medical imaging report generation, organized into four sections: data set, architecture, application, and evaluation. The focus is on frameworks, such as the hierarchical RNN-based framework, attention-based framework, reinforcement learning-based framework, and related works. The paper also discusses potential challenges and future directions for further studies in this area. With the analyzed potential directions for deep learning-based report generation, there are vast opportunities for developments in research and clinical applications. To

gain a more specific understanding of the automatic diagnostic report generation procedure, we plan to conduct further studies on private data sets. Specifically, we aim to establish a radiomics-reporting network to improve the interpretability of deep learning and propose text attention to enhance the readability of medical reports.

Author contributions
TP and PL wrote the main manuscript text and LZ prepared figures and tables. All authors read and approved the final manuscript.

Funding
This work was supported by the Key Scientific Research Project of Universities in Henan Province (China) (23A413002).

Availability of data and materials
Not applicable.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 19 February 2023 Accepted: 9 May 2023
Published online: 18 May 2023

References

1. Monshi MMA, Poon J, Chung V. Deep learning in generating radiology reports: a survey. Artif Intell Med. 2020;106:101878.
2. Hossain MD, Sohel F, Shiratuddin MF, Laga H. A comprehensive survey of deep learning for image captioning. ACM Comput Surveys (CSUR). 2019;51(6):118. https://doi.org/10.1145/3295748.
3. Li Y, Liang X, Hu Z, Xing EP. Hybrid retrieval-generation reinforced agent for medical image report generation. In Advances in Neural Information Processing Systems, 2018, pp. 1530–1540.
4. Pavlopoulos J, Kougia V, Androutsopoulos I. A survey on biomedical image captioning. In: Proceedings of the Second Workshop on Shortcomings in Vision and Language, 2019; pp. 26–36.
5. Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3156–3164).
6. Liang HR, Jiang M, Liang RH, Zhao Q. CapVis: toward better understanding of visual-verbal saliency consistency. ACM Trans Intell Syst Technol. 2019;10(1):23.
7. Yao T, Pan Y, Li Y, Qiu Z, Mei T. Boosting image captioning with attributes. In: Proceedings of the IEEE International Conference on Computer Vision 2017, pp. 4894–4902.
8. Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition 2015, pp. 3156–3164. https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Vinyals_Show_and_Tell_2015_CVPR_paper.html.
9. He XD, Deng L. Deep learning for image-to-text generation a technical overview. IEEE Signal Process Mag. 2017;34(6):109–16. https://doi.org/10.1109/MSP.2017.2741510.
10. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–80. https://doi.org/10.1162/neco.1997.9.8.1735.
11. Chung J, Gulcehre C, Cho K, Bengio Y. Gated feedback recurrent neural networks. In: Proceed of International Conference Machine Learning. 2015.
12. Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. A hierarchical approach for generating descriptive image paragraphs. In the IEEE Conference on Computer Vision and Pattern Recognition, 2017.
13. Guo J, Lu S, Cai H, Zhang W, Yu Y, Wang J. Long text generation via adversarial training with leaked information. In: Thirty-Second AAAI Conference on Artificial Intelligence. 2018
14. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Bengio Y. Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning 2015, pp. 2048–2057.
15. Li LH, Tang S, Zhang YD, Deng LX, Tian Q. GLA: global-local attention for image description. IEEE Trans Multimedia. 2018;20(3):726–37.
16. He XW, Yang Y, Shi BG, Bai X. VD-SAN: visual-densely semantic attention network for image caption generation. Neurocomputing. 2019;328:48–55.
17. Huang FR, Zhang XM, Zhao ZH, Li ZJ. Bi-directional spatial-semantic attention networks for image-text matching. IEEE Trans Image Process. 2019;28(4):2008–20.

18. Kaelbling LP, Littman ML, Moore AW. Reinforcement learning: a survey. JAIR. 1996;4:237–85. https://doi.org/10.1613/jair.301.
19. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. Advances in Neural Information Processing Systems, 2014; pp. 2672–2680. http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.
20. Yan S, Wu F, Smith JS, Lu W, Zhang B. Image captioning using adversarial networks and reinforcement learning. In: 2018 24th International Conference on Pattern Recognition (ICPR) 2018, pp. 248–253.
21. Shin HC, Roberts K, Lu L, Demner-Fushman D, Yao J, Summers RM. Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. In: Proceedings of the IEEE conference on computer vision and pattern recognition 2016, pp. 2497–2506.
22. Kiros R, Salakhutdinov R, Zemel RS. Unifying visual-semantic embeddings with multimodal neural language models. 2014. arXiv preprint arXiv:1411.2539.
23. Jia X, Gavves E, Fernando B, Tuytelaars T. Guiding the long-short term memory model for image caption generation. In: IEEE in ternational Conference on Computer Vision, 2016, pp. 2407–2415. https://doi.org/10.1109/ICCV.2015.277.
24. Greff K, Srivastava RK, Koutník J, Steunebrink BR, Schmidhuber J. Lstm: a search space odyssey. 2015. arXiv:1503.04069v2.
25. Qi W, Shen C, Liu L, Dick A, Hengel A. What value do explicit high level concepts have in vision to language problems? In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 203–212.
26. Wei Y, Xia W, Huang J, Ni B, Dong J, Zhao Y, et al. Cnn: single-label to multi-label. 2014. arXiv:1406.5726v3.
27. Dina DF, Kohli MD, Rosenman MB, Shooshan SE, Laritza R, Sameer A, et al. Preparing a collection of radiology examinations for distribution and retrieval. JAMIA. 2015;2:2.
28. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-Ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. IEEE Conf CVPR. 2017;2017:3462–71. https://doi.org/10.1109/CVPR.2017.369.
29. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, et al. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: National Conference on Artificial Intelligence. Association for the Advancement of Artificial Intelligence (AAAI). 2019.
30. Johnson AEW, Pollard TJ, Berkowitz SJ, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data. 2019;6:317. https://doi.org/10.1038/s41597-019-0322-0.
31. Li Y, Liang X, Hu Z, Xing EP. Hybrid retrieval-generation reinforced agent for medical image report generation. Advances in neural information processing systems, 2018:31.
32. Bustos A, Pertusa A, Salinas JM, Iglesia-Vayá MD. PadChest: a large chest x-ray image dataset with multi-label annotated reports. Med Image Anal. 2020;66: 101797.
33. Heath M, Bowyer K, Kopans D, Moore R, Kegelmeyer W. The digital database for screening mammography. In Proceedings of the 5th international workshop on digital mammography. Medical Physics Publishing. 2001, pp. 212-218. https://www3.nd.edu/kwb/Heath_EtAl_IWDM_2000.pdf.
34. Jing B, Xie P, Xing E. On the automatic generation of medical imaging reports. 2017. arXiv preprint arXiv:1711.08195.
35. Harzig P, Chen YY, Chen F, Lienhart R. Addressing data bias problems for chest x-ray image report generation. 2019. arXiv preprint arXiv:1908.02123.
36. Yuan J, Liao H, Luo R, Luo J. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. 2019. arXiv preprint arXiv:1907.09085.
37. Zhang Z, Xie Y, Xing F, McGough M, Yang L. Mdnet: A semantically and visually interpretable medical image diagnosis network. In: Proceedings of the IEEE conference on computer vision and pattern recognition 2017, pp. 6428–6436.
38. Wang X, Peng Y, Lu L, Lu Z, Summers RM. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In Proceedings of the IEEE conference on computer vision and pattern recognition 2018, pp. 9049–9058.
39. Lee H, Kim ST, Ro YM. Generation of multimodal justification using visual word constraint model for explainable computer-aided diagnosis. 2019. arXiv preprint arXiv:1906.03922.
40. Li X, Cao R, Zhu D. Vispi: automatic visual perception and interpretation of chest x-rays. 2019. arXiv preprint arXiv:1906.05190.
41. Jing B, Wang Z, Xing E. Show, describe and conclude: on exploiting the structure information of chest x-ray reports. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics 2019, pp. 6570–6580.
42. Liu G, Hsu TMH, McDermott M, Boag W., Weng WH, Szolovits P, Ghassemi M. Clinically accurate chest x-ray report generation. 2019. arXiv preprint arXiv:1904.02633.
43. Han, Z., Wei, B., Leung, S., Chung, J., & Li, S. (2018, September). Towards automatic report generation in spine radiology using weakly supervised framework. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 185–193).
44. Xue, Y., Xu, T., Long, L. R., Xue, Z., Antani, S., Thoma, G. R., & Huang, X. (2018). Multimodal recurrent model with attention for automated radiology report generation. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 457–466).
45. Zeng XH, Liu BG, Zhou M. Understanding and generating ultrasound image description. J Comput Sci Technol. 2018;33(5):1086–100.
46. Hicks SA, Pogorelov K, Lange TD, Lux M, Jeppsson, M, Randel KR, et al. Comprehensible reasoning and automated reporting of medical examinations based on deep learning analysis. Acm Multimedia Systems Conference, ACM, 2018, (pp.490–493).
47. Qiao P, Zhang Y, Chen D, Xu G. Character-based convolutional grid neural network for breast cancer classification. In: International Conference on Green Informatics. IEEE Computer Society. 2017

48. Zhang Z, Chen P, Sapkota M, Yang L. TandemNet: distilling knowledge from medical images using diagnostic reports as optional semantic references. Cham: Springer; 2017.
49. Ma K, Wu K, Cheng H, Gu C, Xu R, Guan X. A pathology image diagnosis network with visual interpretability and structured diagnostic report. Cham: Springer; 2018.
50. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002, pp. 311–318.
51. Lin CY. Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out: Proceedings of the ACL-04 workshop, Vol. 8. Barcelona, Spain, 2004.
52. Banerjee S, Lavie A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, vol. 29, 2005, pp. 65-72.
53. Vedantam R, Zitnick CL, Parikh D. Cider: consensus-based image description evaluation. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 4566–4575.
54. Tian J, Li C, Shi Z, Xu F. A diagnostic report generator from CT volumes on liver tumor with semi-supervised attention mechanism. In: International Conference on Medical Image Computing and Computer-Assisted Intervention 2018, pp. 702–710.
55. Ma K, Wu K, Cheng H, Gu C, Xu R, Guan X. A pathology image diagnosis network with visual interpretability and structured diagnostic report. In: International Conference on Neural Information Processing 2018, pp. 282–293.
56. Gale, W., Oakden-Rayner, L., Carneiro, G., Bradley, A. P., & Palmer, L. J. (2018). Producing radiologist-quality reports for interpretable artificial intelligence. 2018. arXiv preprint, arXiv:1806.00340.
57. Hasan SA, Ling Y, Liu J, Sreenivasan R, Anand S, Arora TR, Farri O. Attention-based medical caption generation with image modality classification and clinical concept mapping. In: International Conference of the Cross-Language Evaluation Forum for European Languages, 2018; pp. 224–230.
58. Sun L, Wang W, Li J, Lin J. Study on medical image report generation based on improved encoding-decoding method. In: International Conference on Intelligent Computing, 2019, pp. 686–696.
59. Xie X., Xiong Y, Philip SY, Li K, Zhang S, Zhu Y. Attention-based abnormal-aware gusion network for radiology report generation. In: International Conference on Database Systems for Advanced Applications, 2019, pp. 448–452.
60. Yin C, Qian B, Wei J, Li X., Zhang X., Li Y., et al. Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network. In: 2019 IEEE International Conference on Data Mining (ICDM)., 2020, pp. 728–737. https://doi.org/10.1109/ICDM.2019.00083.
61. Pino P, Parra D, Messina P, Besa C, Uribe S. Inspecting state of the art performance and NLP metrics in image-based medical report generation. 2020. arXiv:2011.09257v2.
62. Zeng X, Wen L, Liu B, Qi X. Deep learning for ultrasound image caption generation based on object detection. Neurocomputing. 2020;392:132–41. https://doi.org/10.1016/j.neucom.2018.11.114.
63. Xu W, Qi C, Xu Z, Lukasiewicz T. Reinforced medical report generation with x-linear attention and repetition penalty. 2020. arXiv:2011.07680v1.
64. Singh S, Karimi S, Ho-Shon K, Hamey L. Show, tell and summarise: learning to generate and summarise radiology findings from medical images. Neural Comput Appl. 2021;33(13):7441–65. https://doi.org/10.1007/s00521-021-05943-6.
65. Yang S, Niu J, Wu J, Wang Y, Li Q. Automatic ultrasound image report generation with adaptive multimodal attention mechanism. Neurocomputing. 2020;427(8):40–9. https://doi.org/10.1016/j.neucom.2020.09.084.
66. Najdenkoska I, Zhen X, Worring M, Shao L. Variational topic inference for chest x-ray report generation. 2021. arXiv:2107.07314.
67. Oa A, Rk B, Ae A, Mh B, Af A. Automated radiology report generation using conditioned transformers. Inform Med Unlocked. 2021;2021(24):100557. https://doi.org/10.1016/j.imu.2021.100557.
68. Liu G, Liao Y, Wang F, Zhang B, Zhang L, Liang X, Wan X, Li S, Li Z, Zhang S, Cui S. Medical-vlbert: Medical visual language bert for covid-19 ct report generation with alternate learning. IEEE Trans Neural Netw Learn Syst. 2021;32(9):3786–97.
69. Han Z, Wei B, Xi X, Chen B, Yin Y, Li S. Unifying neural learning and symbolic reasoning for spinal medical report generation. Med Image Anal. 2021;67:101872.
70. Wu F, Yang H, Peng L, Lian Z, Li M, Qu G, Jiang S, Han Y. AGNet: Automatic generation network for skin imaging reports. Comput Biol Med. 2022;141:105037.
71. Chang YC, Hsing YC, Chiu YW, Shih CC, Lin JH, Hsiao SH, Chen CY. Deep multi-objective learning from low-dose CT for automatic lung-RADS report generation. JPM. 2022;12(3):417.

Publisher's Note