Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Artificial Intelligence In Medicine 106 (2020) 101878

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

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

Deep learning in generating radiology reports: A survey

Maram Mahmoud A. Monshi[a,b,*], Josiah Poon[a], Vera Chung[a]

[a] School of Computer Science, University of Sydney, Sydney, Australia
[b] Department of Information Technology, Taif University, Taif, Saudi Arabia

ARTICLE INFO

Keywords:
Convolutional neural network
Deep learning
Natural language processing
Radiology
Recurrent neural network

ABSTRACT

Substantial progress has been made towards implementing automated radiology reporting models based on deep learning (DL). This is due to the introduction of large medical text/image datasets. Generating radiology coherent paragraphs that do more than traditional medical image annotation, or single sentence-based description, has been the subject of recent academic attention. This presents a more practical and challenging application and moves towards bridging visual medical features and radiologist text. So far, the most common approach has been to utilize publicly available datasets and develop DL models that integrate convolutional neural networks (CNN) for image analysis alongside recurrent neural networks (RNN) for natural language processing (NLP) and natural language generation (NLG). This is an area of research that we anticipate will grow in the near future. We focus our investigation on the following critical challenges: understanding radiology text/image structures and datasets, applying DL algorithms (mainly CNN and RNN), generating radiology text, and improving existing DL based models and evaluation metrics. Lastly, we include a critical discussion and future research recommendations. This survey will be useful for researchers interested in DL, particularly those interested in applying DL to radiology reporting.

1. Introduction

The combination of radiology images and text reports has led to research in generating text reports from images. This was inspired by recent work in generating text descriptions of natural images through inter-modal connections between language and visual features [1]. Traditionally, computer-aided detection (CAD) systems interpret medical images automatically to offer an objective diagnosis and assist radiologists [2]. Unlike CAD, DL is able to learn useful features that move beyond the limitations of radiology detection [3]. For example, DL has been applied to mammography to discriminate between breast cancer and microcalcification [4], on ultrasounds to differentiate breast lesions (malignant and benign), and on CT lung scans to classify pulmonary nodules [5]. Researchers [4,5] noted a significant performance increase in DL models over conventional CAD systems. From a radiologist standpoint, DL helps to improve patient safety by offering more accurate diagnoses, obtains additional diagnostic criteria by generating unobservable data from imaging features, and increases efficiency by performing various tasks automatically [6].

The incapability to construct direct multimodal mapping between radiology images and reports that input an image and output a descriptive report is a well-known shortcoming of most automatic diagnosis methods. The discriminative image features hidden in radiology reports can support better diagnostic conclusion inferences instead of specific image labels. Recent research has utilized this semantic information in reports to propose effective image–text modelling.

Several recent surveys of DL applications [7,8] have been published in healthcare [9], electronic health records (EHR) [10], health informatics [11], medical image analysis [12,13], medicine [14,15], and even radiology [3,6,16,17]. However, no existing reviews specifically address image and text analysis, let alone in radiology. As such, this is the investigative scope of this survey. Papers that cover a wide range of radiology applications and tasks based on DL were analyzed. We found that literature related to generating radiology reports using DL, however, is rare.

In this paper, we examined the DL approaches employed in radiology reporting systems. Unlike other recent surveys that investigated DL in broad health informatics practices ranging from medicine to electronic health records (EHR), our survey focused exclusively on DL techniques tailored to radiology report generation.

2. Radiology

Radiology is a branch of medicine that can be divided into the

* Corresponding author at: School of Computer Science, University of Sydney, Sydney, Australia.
E-mail address: mmon4544@uni.sydney.edu.au (M.M.A. Monshi).

https://doi.org/10.1016/j.artmed.2020.101878
Received 16 April 2019; Received in revised form 30 April 2020; Accepted 10 May 2020

COMPARISON: None
INDICATION: Fatigue, weakness, anterior chest pain
FINDINGS: Cardiomediastinal silhouette and pulmonary vasculature are within normal limits. Lungs are clear. No pneumothorax or pleural effusion. No acute osseous findings.
IMPRESSION: No acute cardiopulmonary findings.

Fig. 1. Example of a radiology report and associated images (obtained from an IU X-ray) [21].

following two subcategories: diagnostic and interventional radiology [18]. Diagnostic radiologists examine medical images to diagnose the cause of a patient's symptoms, monitor treatment effects, screen for various illnesses, and then write radiology reports. On the other hand, interventional radiologists utilize radiology images to guide procedures. Currently, radiology images are interpreted by radiologists who are limited by speed, fatigue, and experience. Certified radiologists are rare due to training costs. As a result, many health-care systems outsource the task of medical image analysis. For example, there are many teleradiology companies in India [12]. Delays or errors in diagnosis can cause harm to patients. Therefore, one solution is for radiology reporting to be performed by an automated, accurate, and efficient DL algorithm.

2.1. Understanding radiology text

A radiology report is a text-based document written by a certified radiologist. It contains descriptive information about a patient's history, symptoms, and interpretations of relevant radiology images [19]. Normally, these reports are written in a specific radiology reporting format and divided into the following sections: comparison, indication, findings, and impressions. The findings section is the most crucial part of the report as it describes medical observations of normal/abnormal features in a presumptive order [20]. Fig. 1 shows an example in the form of an IU X-ray [21] dataset. Here, each report is associated with two chest X-ray images.

A generated radiologist report must follow critical protocols including the correct use of medical terms to describe normal/abnormal diagnoses. They must also include supporting visual evidence in the form of detected disease location and key attributes of the image. There are several lexicons utilized in writing radiology reports including Metathesaurus[1] [22], RadLex[2] [23], and medical subject headings (MeSH).[3] Metathesaurus [22] is a collection of more than five million concept names and a million biomedical terms from over one-hundred controlled vocabulary systems. In contrast, RadLex contains more radiology-specific terms than Metathesaurus including imaging methods and equipment. Furthermore, MeSH offers comprehensive controlled vocabulary created by the United States National Library of Medicine (NLM) to index scientific journal articles and books. For example [24], utilized MeSH terms to mine reports in IU X-rays [21]. However, brain tumors and lung diseases do not have a fixed standardized lexicon. Instead, they have a semi-standardized description system.

The use of DL has shown promising results in generating radiology reports from images [20,25–27]. First, researchers generated a short descriptive sentence of a radiology image using only the image features. Then, they attempted to produce more informative reports with multiple sentences. However, this introduced new challenges in content selection and ordering. Using this method, radiology reports could include information that cannot be detected from image features, such as the nationality of the patient [24]. On the other hand, this text-based DL algorithm is insufficient as it does not include specific image labels.

2.2. Understanding radiology images

There are different types of radiology images, including X-ray, computed tomography (CT), magnetic-resonance imaging (MRI), positron emission tomography (PET), and ultrasound (US) [28]. Fig. 2 shows an example of various radiology imaging modalities and characteristics. Globally, chest radiography is the most common imaging examination that demands correct and immediate interpretation to avoid life-threatening diseases [29]. A single radiologist may need to read and report more than 100 chest X-rays per day [30]. This imaging technology is starting to be employed as the first-line imaging modality by hospitals in Italy and UK to diagnose patients with the coronavirus disease 2019 (COVID-19) [31]. Although chest X-ray is less sensitive than chest CT, it is easy to document and may reduce the risk of cross-infection by utilizing portable radiology units [32]. Recently, several large chest x-rays datasets were released to enable researchers to advance the state-of-the-art for the proposed DL models [29,33]. Consequently, chest X-rays have gained significant attention from DL researchers.

Picture archiving and communication systems (PACS) have been used since the 1990s by modern hospitals for radiology storage, management, transmission, and processing. To enhance standards, digital imaging and communications in medicine (DICOM) was introduced in 1993. It included advanced report and result features [41]. Where DICOM has assisted with many image processing procedures, PACS is an e-system mainly used for the acquisition of medical images.

From DL perspective, radiology images are pre-processed differently due to the varied processor and memory restrictions. Some images, such as X-rays, are two-dimensional (2D) while others such as CT and MRI scans are three-dimensional (3D). Currently, DL models that are trained on simple 2D images are more successful than 3D images which add an extra dimension to the problem [42]. However, experience needs to be gained in applying DL to X-rays because they are 2D projections of a 3D human body [43]. In other words, DL algorithms may need to be adjusted to handle the physiological structures that lie on top of each other in the X-rays. Significantly, DL, in particular CNN, can process an input of 2D and 3D images with only minor adjustments. After all, deep learning in radiology images is still an area of active ongoing research.

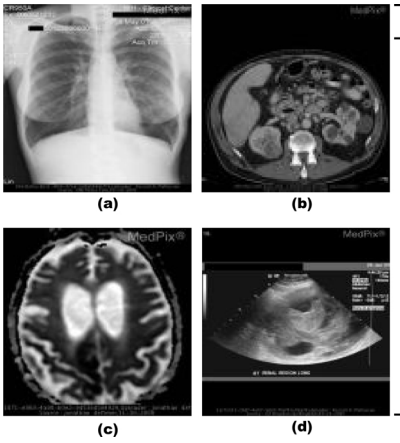
So far, DL has been successfully applied to medical image analysis and acknowledged as a powerful tool for image classification [44], lesion detection [45], segmentation [46], content-based image retrieval (CBIR) [47], report generation from images, and image generation and enhancement [48]. To allow practitioners to rapidly implement DL solutions for image analysis tasks, NiftyNet[4] [49] features an open source framework for many medical imaging CNN algorithms under the

[1] https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus.

[2] https://www.rsna.org/en/practice-tools/data-tools-and-standards/radlex-radiology-lexicon.

[3] https://www.ncbi.nlm.nih.gov/pubmed/.

[4] http://www.niftynet.io.



Modaility	Oragn	Size
X-Ray	Chest	10 MB
CT	Abdomen	250 MB
	Brain	150 MB
	Heart	1 GB
MRI	Abdomen	50 MB
	Brain	100 MB
	Heart	240 MB
PET	Brain	6 MB
	Heart	1 MB
US	Heart	38 MB/s

(e)

Fig. 2. Radiology imaging modalities and characteristics. Note: X-ray (a), CT (b), MRI (c), US (d), image characteristics (e).

Apache License. Several surveys have introduced the role of DL algorithms in medical image analysis, focusing on CNN [12,13]. Biswas et al. [50] classifies DL models based on application area, including cardiovascular, neurology, mammography, microscopy, dermatology, gastroenterology, and pulmonary applications.

2.3. Text/Image radiology dataset

Table 1 compares publicly available radiology image datasets with relevant reports in the medical informatics domain. These include the following: the Indiana University chest X-ray (IU X-ray) [21], ChestX-ray14 [34], MIMIC-CXR [33], pathology detection in chest radiographs (PadChest) [37], the digital database for screening mammography (DDSM), and the pathology education informational resource (PEIR). Researchers have employed these multimodal medical databases for developing and evaluating DL models. Nevertheless, there are few large and accessible datasets adequate for developing CNN models. In addition, researchers conduct experiments using different database subsets. This makes it difficult to compare the performance of their proposed approaches.

At present, IU X-ray [21] and ChestX-ray14 [34] are the most frequently used datasets by researchers in the medical informatics domain. The IU X-ray [21] collection consists of 7470 chest X-rays with 3955 radiology reports available through OpenI. OpenI is an open-source collection of literature and biomedical images. It contains IU X-ray, 2064 orthopedic illustrations, and more than three million images from PubMed and the National Library of Medicine (NLM). Researchers [20,24–27] have used this dataset to demonstrate how their proposed DL models label and describe the diseases associated with the images. However, data in IU X-ray comes from fully anonymized reports in two hospitals. As a result, some keywords, findings and images are missing. ChestX-ray14 [34] is from the national institute of health (NIH) clinical center. It is an open access chest X-ray dataset that includes 112,120 X-ray images with fourteen thorax disease labels (atelectasis, consolidation, infiltration, pneumothorax, edema, emphysema, fibrosis, effusion, pneumonia, pleural thickening, cardiomegaly, nodule, mass, and hernia). These labels were mined from the original radiologist reports. However, the complete text reports are not publicly available.

CheXpert [29] and MIMIC-CXR [33] are the latest co-released open source datasets that use the CheXpert labeler to extract annotations from unstructured radiology reports. CheXpert is a dataset that consists of 224,316 chest radiographs from 65,240 patients labeled due to the presence of 14 common chest radiographic observations. ChestX-ray14 uses an automatic labeler to extract labels from reports. On the other hand, CheXpert offers radiologists labeled validation and expert scores. The largest open access chest radiography to date is MIMIC-CXR. This includes 371,920 chest X-rays linked to 227,943 reports gathered from the Beth Israel Deaconess Medical Center. Through a limited release version of this dataset [35], conducted the first work that trained a collection of CNNs using a huge dataset to recognize thorax diseases. Then [36], used MIMIC-CXR v1.0.0. to show that processing multi-view chest X-rays simultaneously resulted in better classification performance.

PadChest [37], however, is labeled with the largest number of annotations including 174 radiology findings, 19 diagnoses, and 104 anatomic locations. This dataset contains 160,868 chest X-rays from six different views and the associated 109,931 reports collected from San Juan Hospital. It provides researchers with the opportunity to address unfinished investigations such as measuring DL model performance using the chest X-ray views [38].

Apart from X-ray collections, DDSM [39] and PEIR are open source

Table 1
Radiology image/text dataset (available online).

Dataset	Description	Base annotation	Employed by
IU X-Ray[1] Demner-Fushman, et al. [21] 2015	7470 chest x-rays 3955 radiology reports	Thorax diseases	[20,24–27]
ChestX-ray14[2] Wang, et al. [34] 2017	112,120 chest x-rays 14 thoracic labels	Atelectasis, consolidation, infiltration, pneumothorax, edema, emphysema, fibrosis, effusion, pneumonia, pleural thickening, cardiomegaly, nodule, mass and hernia	[20,27,29]
CheXpert[3] Irvin, et al. [29] 2019	224,316 chest x-rays 14 annotated observations	No finding, enlarged cardamom, cardiomegaly, lung opacity, lung lesion, edema, consolidation, pneumonia, atelectasis, pneumothorax, pleural effusion, pleural other, fracture, support devices	–
MIMIC-CXR[4] Johnson, et al. [33] 2019	371,920 chest x-rays 227,943 studies		[35,36]
PadChest[5] Bustos, et al. [37] 2019	160,868 chest x-rays 109.931 Spanish reports	174 radiology findings, 19 diagnoses and 104 anatomic locations	[38]
PEIR Digital Library[6]	4732 images in 20 categories one sentence per image	Multiple (e.g. abdomen, adrenal, aorta, breast, chest, heads and kidney)	[25]
DDSM[7] Heath, et al. [39] 2000	2620 breast mammography 3 labels	Normal, benign and malignant	[40]

[1] https://openi.nlm.nih.gov/faq.php.
[2] https://nihcc.app.box.com/v/ChestXray-NIHCC.
[3] https://stanfordmlgroup.github.io/competitions/chexpert/.
[4] https://archive.physionet.org/physiobank/database/mimiccxr/.
[5] http://bimcv.cipf.es/bimcv-projects/padchest/.
[6] http://peir.path.uab.edu/library/index.php?/category/106.
[7] http://marathon.csee.usf.edu/Mammography/Database.html.

datasets of different image modality. For example, PEIR is a digital library 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, and kidneys. On the other hand, DDSM [39] 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. Kisilev et al. [40] selected a subset of the DDSM database that consisted of 974 images annotated with semantic descriptors to test their multi-task-loss CNN based model. This outperforms the accuracy of current techniques by up to 10% when detecting and describing lesions.

Moreover, researchers have trained their deep learning frameworks on several privately-owned datasets, including the PACS from the NIH clinical center [51] and CX−CHR [62]. The PACS from the NIH clinical center consists of 216,000 2D images with radiology reports that offer visual references to pathologies. The CX−CHR dataset contains chest X-rays of 35,500 patients and contains Chinese reports.

3. Deep learning (DL)

Currently, DL is a promising subfield of machine learning (ML) which, in turn, is a subfield of artificial intelligence (AI) (Fig. 3a). Artificial intelligence occurs when a machine is composed of multiple layers, uses raw data as input, and improves the representations required for pattern recognition [52]. Essentially, a linear combination v_k of input signals x_1, x_2, x_3, ... x_m adds bias b_k to apply an affine transformation and generate the output y_k (Fig. 3b) where w_{k1}, w_{k2}, w_{k3}, ... w_{km} are the weights, and $\varphi(.)$ is the activation function (described in Section 3.1). This main computational element, known as the neuron or perceptron, enables the DL machine to learn from experience without the need to specify the desired knowledge. Currently, DL has already succeeded in many computerized applications including computer vision, NLP, speech processing, gaming, and cross-media retrieval. From a radiology perspective, DL models can be fed with multiple datatypes and iteratively distort them as they flow from layer to layer [9] (Fig. 3c). This is a particularly relevant function for radiology data as it consists of reports and linked images.

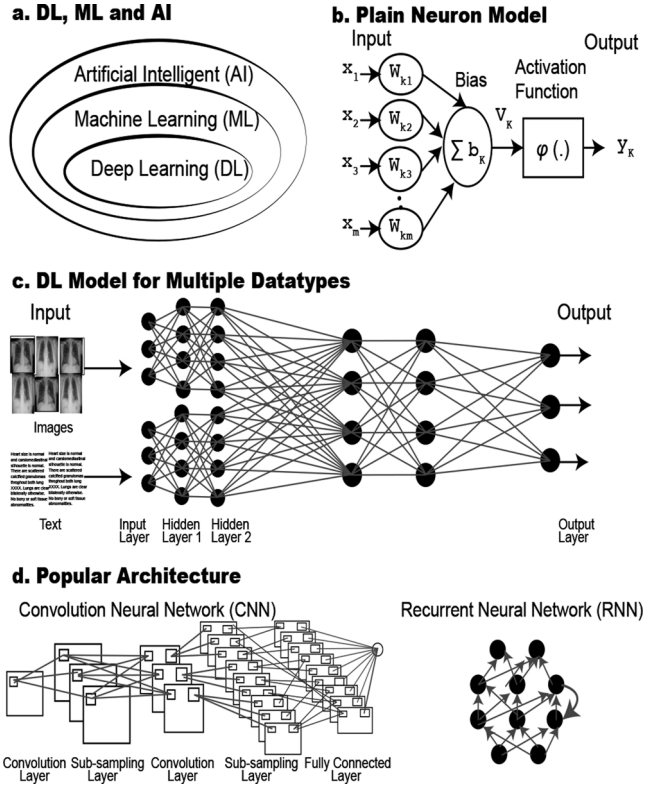


Fig. 3. Deep learning.

Researchers have classified DL models into three categories: supervised, unsupervised, and reinforcement learning (RL) [8,10]. Supervised learning mainly infers a mapping function $y = f(x)$ from input x to output y such as multilayer perceptron (MLP), recurrent neural network (RNN), and convolutional neural network (CNN). Often RNNs are accompanied with CNNs to generate medical image descriptions [24,27,51,53] (Fig. 3d). In contrast, unsupervised DL takes onboard remarkable properties related to the distribution of x including Boltzmann machines (BM) and autoencoders (AE). Deep RL is a semi-supervised technique for partially labeled datasets as it can act with limited input data. For instance, if a deep RL network is fed with several tumor cells, it can overinterpret an image to detect insignificant aspects [54]. To enable effective and robust radiology report generation, using RL, HRGR-Agent [20] trained the retrieval policy module and the generation module using sentence-level and word-level rewards, respectively.

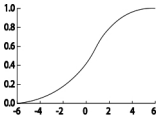
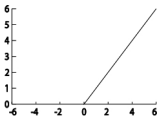
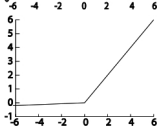
3.1. Activation function

An activation function is a critical element of DL as it adds nonlinearity by taking the weighted sum of inputs in one layer and converting it into an output value [16]. Then, this value is conveyed to nodes in the subsequent layer. Table 2 illustrates common activation functions including sigmoidal, hyperbolic tangent (TanH), rectified linear unit (ReLU) [55], and leaky ReLU [56]. Sigmoidal is one of the earliest activation methods used in neural networks but can cause network instability or freeze network learning. The limitations of TanH are similar as it is a scaled form of the sigmoid function.

On the other hand, ReLU performs better than sigmoidal functions as it was the first to be successfully used for neural networks by [55]. It converts the weighted sum of inputs to zero if they are less than zero or to the same input if they are equal to or greater than zero. Leaky ReLU is an extension of ReLU that outputs small negative numbers if the inputs are negative. If not, it produces the same outputs as ReLU. Researchers tend to begin with ReLU and then apply other activation functions if they do not obtain optimal results.

All traditional CNN activation functions output a single result for a single input except Softmax. Instead, Softmax produces multiple outputs. It is useful as it converts the output of the last neural network layer into a probability distribution. In practice, Softmax is used in

Table 2
Activation function for DL.

Name	Equation	Plot	Characteristics
Sigmoid	$sigmoid\ (x) = \frac{1}{(1+e^{-x})}$		Range [0,[1] Not zero centered Have exponential centered
TanH	$tahn(x) = \frac{2}{(1+e^{-2x})} - 1$		Range [-1, 1] Zero centered
ReLU [55]	$ReLU\ (x) = 0,\ x < 0$ OR $= x,\ x \geq 0$		It doesn't saturate Fast
leaky ReLU [56]	$leaky\ ReLU\ (x) = x,\ x < 0$ OR $= \partial x,\ x \geq 0$		Overcome dead ReLU problem

multiclass classifications, while sigmoid is used in binary classifications [57].

3.2. Convolutional neural network (CNN)

A CNN [58] is a type of multi-layer neural network that uses minimal processing to recognize visual patterns from pixel images. One of the main advantages of CNN is its ability to automatically amalgamate low-level features (including lines and edges) into high-level features (such as shapes) within subsequent layers [12]. For each convolutional layer l, a set of k kernels W_1, W_2, ..., W_k with biases b_1, b_2, ..., b_k convolve an input image to generate feature maps X_k. These generated maps have a non-linear transform $\varphi(.)$ in each layer (refer to Eq. 1.).

$$X_k^l = \varphi\left(W_k^{l-1} * X^{l-1} + b_k^{l-1}\right) \quad (1)$$

There are several CNN models including deep feed-forward CNNs for images and word-embedding networks for text. The histogram of oriented gradients (HOG) and scale-invariant feature transform (SIFT) are two examples of convolutional image features. However, deep CNNs significantly outperform shallow learning frameworks and hand-crafted image features as they need larger collections of training data [59].

Recently, CNNs have become the primary frameworks for mining medical data as the number of papers published on CNN methods and applications has increased rapidly since 2015 [12,13]. In radiology, CNN is the most applicable DL algorithm for performing various tasks including medical image classification and segmentation [60]. Interestingly, CNNs can transfer learning from a large database unrelated to the current task (e.g., ImageNet) into a related one (e.g., IU X-ray).

3.2.1. Architecture

The most popular CNN architectures were proposed by top competitors at the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). This includes the following architectures: AlexNet [61], ZFNet [62], Visual Geometry Group (VGG-16) [63], GoogLeNet [64], Residual Network (ResNet) [65], ResNeXt [66], CUImage Team [67], and SENets [68] (see Table 3). ImageNet is a project that aims to create an enormous visual database that can be utilized by researchers in the field of visual object recognition [69]. It should be noted that ImageNet runs ILSVRC, an annual contest where software programmers classify and detect objects and scenes.

In 2012, [61] noted how AlexNet was the first model to considerably improve image classification performance. It obtained a 16.4 % error rate using the ImageNet dataset. This model minimized the overfitting problem using data augmentation and dropout procedures. Two remarkable models were then proposed in 2014: the VGG-16 (7.4 % error rate), which reduced the spatial size of the input in each layer, and GoogLeNet (6.67 % error rate), which permitted procedures such as pooling and convolutional to run in parallel to each other. AlexNet uses eight convolutional layers, 650,000 neurons (60,000,000 parameters) and has an error rate of 16.4 %. In contrast, VGG-16 consist of 16 convolutional layers, 133,000,000 parameters and 7.4 % error rates [70]. It is clear that VGG-16 is a significantly deeper model than AlexNet, which is why its error rate is lower.

Table 3
CNN architectures (ILSVRC winners).

Winer by year	No. of conv. layers	Top-5 error rate (%)
2012 - AlexNet [61]	8	16.4
2013 - ZFNet [62]	8	11.7
2014 second - VGG-16 [63]	16	7.4
2014 first - GoogLeNet [64]	22	6.67
2015 - ResNet [65]	152	3.57
2016 second – ResNeXt [66]	101	3.03
2016 first – CUImage Team [67]	152	2.99
2017 - SENets [68]	152	2.25

By 2015, automatic image classification models could outperform human manual annotation with a 5 %–10 % error, respectively. This first occurred when [65] introduced Microsoft deep ResNet. This contains 152 layers that apply residual connections in CNNs to address the issues of vanishing gradients [71] and degradation. The ILSVRC 2016 winner was the CUImage team [67], who assembled the following six architectures: Inception v3, Inception v4, Inception ResNet v2, ResNet 200, Wide Resnet 68, and Wide Resnet 3. However, the 2016 runner-up, ResNext [66], introduced a simple framework that consisted of branches in a residual block. Each branch conducted a transformation aggregated by a summation function at the end. Although this model is based on ResNet and uses less layers, it outperforms ResNet, Inception-v3 and Inception-ResNet [72]. It can be generalizable by reshaping it using other models like AlexNet.

In 2017, the ILSVRC concluded as researchers considered the problem of supervised image classification solved [7]. The 2017 winner was squeeze and excitation networks (SENet). This network is based on the ResNeXt-152 model and adds recalibration to adaptively reweight feature maps.

To generate radiology reports, researchers follow some ImageNet CNN network settings as well as other reliable architectures. These include network in network (NIN) [73] and densely connected convolutional network (DenseNet) [74] with slight modifications. For instance [24], notes that AlexNet is a complex method. Instead, they use NIN as it is a simpler and faster model. In addition, they suggest that GoogLeNet is the baseline CNN model and use it to train their data. Although AlexNet and GoogLeNet have different depths, Wang et al. [59] utilized both to train their looped deep pseudo-task optimization network model (LDPO). When extracting features from images, VGG16 is the preferred choice for the majority of researches in the visual pattern recognition community [19]. This is largely because VGG16 offers a uniform CNN architecture and publicly available weight configuration[5]. For example, [51,53] adopt this architecture to read radiology images.

3.3. Recurrent neural network (RNN)

RNN is a neural network that processes sequential information while maintaining a state vector within its hidden neurons [75]. Eq. (2) is the basic RNN that preserves a hidden state h at a time t that is the outcome of a non-linear mapping sing its input x_t and the previous state h_{t-1}, where W and R are the shared weight matrices over time. On the other hand, CNNs are the preferable networks for pixels in an image and other clear spatial structure data. Recurrent neural networks work well with natural language and similar sequentially ordered data [10]. They can predict next words based on the former ones in the language model [76]. However, it is hard to save information for a long time as the weights are equal in all RNN layers. Another issue is the requirement for a backpropagation algorithm to train RNN as the gradients either grow or shrink. Consequently, variations of RNN have been introduced to overcome these limitations.

$$h_t = \varphi\left(W_{x_t} + Rh_{t-1} + b\right) \quad (2)$$

The most popular extensions of RNN are Long Short-Term Memory (LSTM) [77] and the Gated Recurrent Unit (GRU) [78]. Long short-term memory uses memory blocks to save the network temporal state and gates to monitor the information flow. On the other hand, GRU is a lighter form of RNN than LSTM in terms of topology, computation expenses, and complexity. At present, researchers must choose between the faster model offered by GRU that needs fewer parameters or the higher performing model provided by LSTM that contains sufficient data and computational power [8].

[5] http://www.robots.ox.ac.uk/~vgg/research/very_deep/.

3.4. Software

Convolutional architecture for fast feature embedding (Caffe)[6] [79] is the most common software package utilized by practitioners to automate radiology reporting. Using Caffe [51], trained their deep CNN model to map X-rays into specified document categories, and [40] implemented a multi task loss CNN model to describe medical images. Using Caffe [53,59,80], acquired pre-trained CNN models on ImageNet for their radiology annotation systems.

However, there are several other software packages that support CNN and RNN implementations, including TensorFlow[7] [81] and PyTorch[8] [82]. Using both TensorFlow and Tensorpack[9], [27] implement a text–image embedding network (TieNet) that produces thorax diseases reports. DualNet [35] and the hybrid retrieval-generation reinforced agent (HRGR-Agent) [20] frameworks are based on PyTorch. These software packages are open-source projects that utilize Nvidia support to enhance performance through graphics processing unit (GPU) acceleration. To note, training DL can be accelerated through advanced GPU that facilitates parallel processing.

4. Generating radiology text

Natural language processing (NLP) explores the use of machines to process/understand human languages and carry out useful tasks. Traditional learning algorithms for NLP are often incapable of absorbing a large volumes of training data as feature engineering requires significant human expertise [83]. Several years ago, NLP was brought forward by a new era of deep learning algorithms using a vision named "NLP from scratch" [84]. Such DL waves have the capacity to learn representations from text through layers of nonlinear neurons for feature extraction. Since 2010, DL has been productively applied to NLP tasks [85] including natural language generation (NLG) from meaning representation. This can be considered the inverse of natural language understanding [86]. Through this, DL can generate fluent, communicative, and new image descriptions.

Applied to a free-form radiologist text, NLP assists with converting text into a structured report, extracting meaningful information, and classifying reports [87]. A recent NLP technique is neural language modelling, which includes word embedding and recurrent language models [88]. Word embedding converts words into vectors to allow less sparse data representation. Using this, DL models can be trained with smaller datasets. Advanced word embedding was applied to a large collection of radiology reports to generate word vectors of radiology image descriptions [20,25–27,51,89]. Recurrent language models predict word output based on a sequence of arbitrary past words. As such, they are not limited by fixed input dimensions.

Generally, radiology reports are semi-structured and use standardized documentation templates [33]. Consequently, researchers have proposed open-source NLP tools to extract controlled vocabulary from radiology reports. Examples of these tools include NegBio labeler[10] [28] and CheXpert labeler[11]. NegBio was developed by NIH and used to annotate the ChestX-ray14 dataset. CheXpert was built by the Stanford Machine Learning Group and based on NegBio. However, CheXpert achieved a higher F1 score.

5. DL models for generating radiology report

Overall, the purpose of the proposed models was to generate interpretations of radiology images. During training, the input for these models was a collection of images and associated reports, as shown in Fig. 4. First, researchers proposed models to align disease descriptions to the relevant visual regions using multimodal embedding. They then used the outcomes as training data for additional models. This training data allowed the additional models to learn how to generate the image descriptions.

Table 4 categorizes the existing approaches into three main levels to summarize their main characteristics. These categories are as follows: words, sentences, and paragraphs. It is clear that the accessibility of a large volume of radiology reports and images allowed deep CNNs to become the premier learning method and address the automatic text report generation issue.

Table 5 compares the results of the generated reports through quantitative evaluation matrices (defined in section 6.1). To the best of our knowledge, the multi-task learning model [25] outperforms existing approaches in generating radiology paragraphs using the IU X-ray dataset.

5.1. Word level

In 2015, the first text/image DL framework with a large-scale PACS was proposed by [51] and used in a national research hospital. This process is explained in more detail in [19]. This system uses approximately 780,000 radiology reports and around 216,000 2D images to extract and mine the semantic interactions between them. This framework is capable of matching images with their descriptions automatically using NLP. Latent Dirichlet Allocation (LDA) [90] was applied to obtain the semantic interpretation of diagnostic images, and a CNN was trained to map the images into document categories. The weak supervision method was used to generate interpretations of radiology images, and the strict supervision method was used to detect the absence or presence of several common diseases. In the testing set, the match rate between predicted disease words and actual words in the report was 0.56. This system represents a significant step towards accurately generating radiologist reports using enormous medical image databases.

Nevertheless, the clusters in [51] are highly unbalanced. This is because most images are clustered into three groups as they were derived from text modalities only (approximately 780,000 reports). On the other hand, Wang et al. [59] created the LDPO model, which formed clusters from text reports as well as image cues to offer a more visually coherent and balanced method in terms of clusters. As such, LDPO is an iterative system that extracts deep CNN features based on fine-tuned radiologist topic labels and mutual information shared between discovered clusters. Afterwards, the framework either stops the iteration and outputs optimized clustering or inputs the refined cluster labels into the next iteration to fine-tune the CNN model. At the end, NLP is applied to the radiology reports to count and rank the frequency of each word. This process allocates the most common words, which are then used as the keyword labels for each cluster. To evaluate the system, a board of certified radiologists reviewed the resultant keywords and sampled images. The results of applying the LDPO model to discovery clusters were found to be visually coherent and highly balanced clusters. Nevertheless, the looped property is specific to deep CNN classification-clustering methods as other kinds of classifiers cannot learn satisfactory image characteristics simultaneously.

Using a dataset of more than 16,000 X-ray images and Chinese radiology reports, [53] trained a CNN model to automatically label new images with one of ten pre-defined labels: normal, increased lung marking, aortosclerosis, increased heart shadow, pleural thickening, pulmonary interstitial hyperplasia, costophrenic angle blunting, pleural effusion, emphysema, and bronchitis. These disease labels were extracted from the reports using basic NLP techniques. In addition, this system can generate the correct label with an accuracy of 97 %. However, it performed poorly in cases including increased heart shadows

[6] http://caffe.berkeleyvision.org/.
[7] https://www.tensorflow.org/.
[8] http://pytorch.org/.
[9] https://github.com/ppwwyyxx/tensorpack/.
[10] https://github.com/ncbi-nlp/NegBio.
[11] https://github.com/stanfordmlgroup/chexpert-labeler.

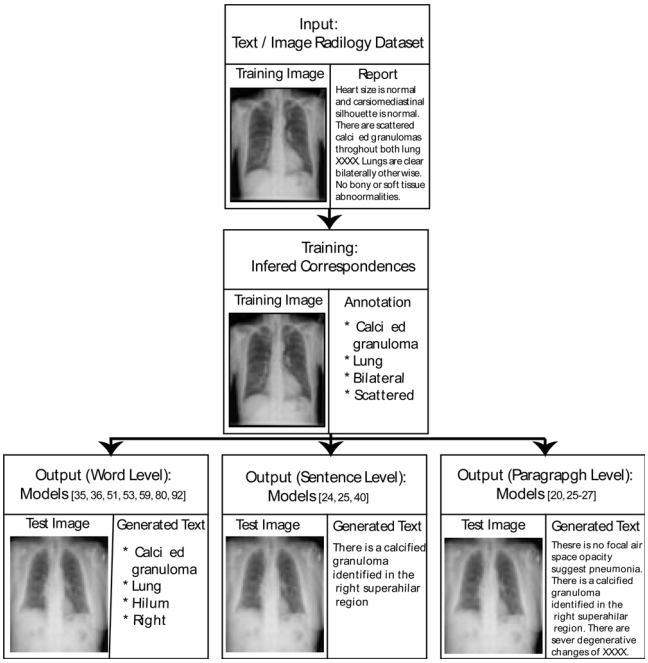


Fig. 4. Framework of the radiology reporting models.

and pleural thickening due to the unbalanced database. In this dataset, half of the images were labelled as "normal" cases.

The above frameworks involve two separate models. Therefore, a single model trained end-to-end that can move directly from a radiology text-image database to region-level annotation has yet to be created.

CheXNet [80] is one of the most popular DL models that utilized the Chest-Xray14 dataset [34]. It contains more than 112,000 images from a reformed version of DenseNet with 121 convolution layers. CheXNet outperformed a panel of three radiologists when annotating pneumonia and 13 other diseases. Furthermore, it applied class activation mapping (CAM) [91] to produce heatmaps that visualized the indicative regions of the disease in the image. Using the same dataset but with ResNet-152 architecture instead, ChestNet [92] incorporated an additional attention branch into CNN based on gradient-weighted class activation mapping (Grad-CAM) [93]. This exploited the correlation between labels and disease locations.

DualNet [35] and the multi-view model [36] employed the MIMIC-CXR [33] dataset, which is over four times the size of Chest-Xray14 [34], to demonstrate the benefits of simultaneously processing frontal and lateral chest X-rays when detecting common thorax diseases. They used DenseNet-121 and ResNet-50, respectively. The multi-view model adopted discriminative learning rates [94] and introduced the stage wise training approach to reduce training time and increase accuracy. This had an average labelling performance of 0.779 AUC.

5.2. Sentence level

In contrast to recent studies that only detected diseases in images using text/image datasets [35,36,51,53,59,80,92]. Shin et al. [24] described the context of the disease in a similar way to a radiology report. They introduced a recurrent neural cascade model to detect and describe disease location, severity, and the affected organs to offer a better understanding of the disease. This system computed labels based on joint text/image contexts after initial CNN/RNN training using single object labels in a chest X-ray dataset from IU X-ray [21]. Eventually, it generated image descriptions by training the RNN with the new CNN image embedding (refer to Eq. 3.), where I denotes the input image, t is the time step, N is the number of words in the annotation, Y is the output word, S is the correct word and $h_{im:text}$ represents the joint image/text context vector from the first iteration, $iter = 0$.

$$L(I,\ S) = -\sum_{t=1}^{N} [P_{RNN_{iter=1}}(y_t = S_t) \mid \{CNN_{iter=1}(I) \mid h_{im:text_{iter=0}}\}] \tag{3}$$

Similarly, the multi-task-loss CNN-based system generated radiologist sentences to describe tumor lesions (shape, margin, and density) in breast images [40]. Essentially, this system was trained using a DDSM dataset and a private dataset of mammography and ultrasound to produce and rank the rectangular regions of interest (ROIs). The highest ROIs were fed into the remaining network layers which, in turn, generated semantic descriptions of subsequent ROIs. This system provided automatic lesion detection in breast images alongside semantic descriptions. Jing et al. [25] added a co-attention mechanism to describe abnormal lesions by discovering visual and semantic information.

5.3. Paragraph level

The first work towards generating truly radiology reports with long and diverse topics is a multitask learning model with a co-attention mechanism. It contains a hierarchical LSTM to produce long descriptive paragraphs through capturing long-range semantics [25]. Although this model achieved outstanding results when generating descriptive radiology reports using the IU X-ray dataset, the produced paragraphs contained repeated sentences due to a lack of contextual coherence in the hierarchical models.

On the other hand, [26] generated sentences using the same dataset through an attention input of image encoding and the first generated sentence. This method maintained coherence in the resultant paragraphs as it uses CNN and LSTM in a recurrent way. As [26] filtered reports without two associated images (frontal and lateral chest X-rays) and reports without complete sections from the IU X-ray dataset, the training was performed using a small dataset. As a result, the generated text was missing some abnormal descriptions and contained sentences that were different from the ones in the training set.

Using the same dataset, [27] proposed a text-image embedding network (TieNet) that integrated multi-level attention with a CNN-RNN framework for classification and reporting. The CNN, RNN, and LSTM were based on ResNet-50, the visual spatial attention approach [96], and standard LSTM, respectively. Multiple RNNs may have enhanced TieNet by learning the disease attributes more efficiently which, in turn, may have improved the auto-report quality.

Recently, [20] introduced the first retrieval model with a generative neural network using RL. This is called the hybrid retrieval-generation reinforced agent (HRGR-Agent). The HRGR-Agent extracts visual features of chest X-rays from the last convolutional layer of DenseNet or VGG19 and improves text generation by empowering RNN with an attention mechanism. The experiments on two medical databases, IU X-ray and CX-CHR, showed high performance in generating precise text that described rare abnormal findings. The CX-CHR database utilized was a proprietary dataset of Chinese reports and linked images. This made it difficult to compare the HRGR-Agent with other recent state-of-the-art models.

In contrast, [97] used the largest public intensive care unit (ICU) patient dataset to introduce a framework that learned multiple disease labels from two types of features: medical charts and notes. Instead of considering the correlation between diseases in the same way as existing methods, this approach used disease-specific features. However, the paper only demonstrated an intuitive implementation of the disease-specific feature construction, rather than using multiple clusters for positive and negative instances.

Table 4
DL models for generating radiology report.

Model	Proposed by	Image Modality	Dataset	Organ	Pathology	Software	CNN Architecture	Base	
								Technique	Task
Word-level									
Deep mining model	Shin, et al. [51] 2015	CT MR PET	PACS of NIH clinical centre [62]	Multiple (e.g., neck, bone, liver, brain and heart)	Multiple (e.g. adenopathy, metastasis and sinus diseases)	Caffe [79]	AlexNet [61] VGG-16 [63] VGG-19 [63]	LDA & RNN CNN	Generate semantic labels Map from images to label spaces
LDPO: looped deep pseudo task optimization network	Wang, et al. [59] 2016	Computed radiography Ultrasound				Caffe [79]	AlexNet [61] GoogLeNet [64]	CNN K-means/RIM NLP PCA	Initialize looped optimization Cluster images Extracts semantically relevant words Reduce dimensionality
CNN-based classification model	Dong, et al. [53] 2017	X-Ray	PACS of the fourth people's hospital (Chinees reports)	Chest	9 diseases (e.g. emphysema & bronchitis)	Caffe [79]	VGG-16 [63] ResNet-101 [65]	NLP CNN RNN	Extract disease labels from reports Classify images Describe a detected disease
CheXNet	Rajpurkar, et al. [80] 2017		ChestX-ray14 [34]		Pneumonia & 13 other pathologies	-	DenseNet [74]	CNN CAM [91]	Classify images Produce heatmaps
ChestNet	Wang and Xia [92] 2018					Caffe [79]	Resnet-152 [65]	CNN Attention mechanism (Grad-CAM [93])	Perform feature extraction-classification Exploits correlation between class labels & pathology locations
DualNet	Rubin, et al. [35] 2019		MIMIC-CXR [33]		14 Thorax diseases (e.g pneumonia & edema)	PyTorch [82]	DenseNet-121 [74]	NLP (NegBio [95]) CNN	Map reports into UMLS concept ids Recognize multiple diseases
Multi-view model	Monshi, et al. [36] 2019				12 Thorax diseases		Resnet-50 [65]	CNN discriminative learning rates [94]	Detect diseases Tune each layer with various learning rates
Sentence-level									
Recurrent neural cascade model	Shin, et al. [24] 2016	X-Ray	IU X-Ray [21]	Chest	Thorax diseases (e.g. cardiomegaly, and granuloma)	-	NIN [73] GoogLeNet [64]	CNN LSTM-RNN [77] / GRU-RNN [78]	Classify images Describe disease contexts
Multi-task-loss CNN model	Kisilev, et al. [40] 2016	Mammograph Ultrasound	DDSM Private dataset [34]	Breast	Tumour	Caffe [79]	AlexNet (5 conv. layers) [61]	CNN	Produce ranked ROI Generate semantic description
Multi-task learning model	Jing, et al. [25] 2017	Multiple	PEIR Gross	21 organ categories (e.g. kidney)	Multiple	-	VGG-19 [63]	CNN MLC	Learn visual features Predict relevant tags
Paragraph-level									

(*continued on next page*)

Table 4 (*continued*)

Model	Proposed by	Image Modality	Dataset	Organ	Pathology	Software	CNN Architecture	Base	
								Technique	Task
Multi-task learning model	Jing, et al. [25] 2017	X-Ray	IU X-Ray [21]	Chest	Thorax diseases	-	VGG-19 [63]	CNN hierarchical LSTM MLC	Learn visual features Generate long paragraphs Predict relevant tags
Multimodal recurrent model with attention	Xue, et al. [26] 2018		IU X-Ray [21]			-	Resnet-152 [65]	CNN Single layer LSTM Bi-LSTM and ID CNN	Extract visual features Sentence decoding Sentence encoding
TieNet: text-image embedding network	Wang, et al. [27] 2018		IU X-Ray [21]			TensorFlow [81] Tensorpack	ResNet-50 [65]	NLP CNN-RNN LSTM-RNN	Mine disease labels Link words with image regions Produce reports
HRGR-Agent: hybrid retrieval-generation reinforced agent	Li, et al. [20] 2018		IU X-Ray [21] CX-CHR (Chinese reports) [20]			PyTorch [82].	DensNet [74] VGG19 [63]	CNN	Extract visual features

6. Evaluation

Evaluating radiology reporting models has become increasingly essential due to the rapid introduction of DL approaches to large medical datasets. Both quantitative (machine-based) and qualitative (human-based) evaluations have been employed to compare the benchmark reporting models. Qualitative evaluation is more expensive than quantitative and is not repeatable. However, it may offer additional valuable measurement for generated reports.

6.1. Quantitative

The common evaluation metrics for image captioning and machine learning are bilingual evaluation understudy (BLEU) [98], recall-oriented understudy for gisting evaluation (ROUGE) [99], METEOR [100], consensus-based image description evaluation (CIDEr) [101], and semantic propositional image caption evaluation (SPICE) [102]. Table 6 compares these matrices using their original purposes, main ideas, strengths, and weaknesses.

These evaluation matrices are employed by researchers to compare their proposed models of generating radiology reports against the benchmarks. They automatically calculate an accuracy score for a new model by observing the similarity/differences between the generated captions and the radiologist's written descriptions from empirical observation. Increased performance is indicated through higher scores in BLEU, ROUGE, METEOR, CIDEr, and SPICE. The MS COCO evaluation kit[12] offers the implementation script for these evaluation matrices in terms of caption generation.

BLEU-n metrices [98] are precision metrices for machine translation that are computed by multiplying n-gram precision scores by a penalty for short sentences. They have been employed to measure the similarity between a pair of sentences. A superior version of BLEU was proposed by [103]. However, BLEU suffers from a low performance in explicit word matching.

ROUGE [99] is a recall metric for summarization systems that matches intersecting n-grams, word sequences, and word pairs. ROUGE-L is a version of ROUGE that calculates the longest common sub sequences between two sentences.

METEOR [100] is a recall metric for machine translation that utilizes synonyms, paraphrase matching, precision, and unigram recall to obtain harmonic overlapping between sentences. It overcomes BLEU's weaknesses in failing to locate semantic similarity by applying synonym matching based on WordNet. Nonetheless, observing synonyms alone may not be adequate to capture semantic similarity.

CIDEr [101] is an evaluation metric for image captioning that calculates cosine similarity between candidate image c_i annotation and the associated sentences produced by humans. It works in a purely linguistic means, but its evaluations are ineffective as it sometimes provides large weight for insignificant sentence details.

SPICE [102] is a recent evaluation metric for image caption that uses scene-graph tuples to parses a sentence into semantic tokens including object classes, relation types, and attribute types. Thus, the quality of the parsing determines CIDEr's performance. In some cases, this may result in failure as illustrated by an example in [104]. In a similar way to METEOR, SPICE utilizes WordNet synonym matching for tuple matching.

The different design choices of evaluation metrics, such as n-gram and scene-graph, result in metrics that have different strengths and weaknesses. For example, BLEU, ROUGE, and CIDEr use only exact n-gram matches, but METEOR adds synonyms and paraphrases. Although BLEU is based on precision, METEOR and ROUGE are recall-based metrics. As a consequence, [104] suggested that existing evaluation metrics should complement each other in measuring the quality,

[12] https://github.com/tylin/coco-caption.

Table 5
Quantitative evaluation of generated radiology reports based on DL models.

Model		Database	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGH	ROUGH_L	CIDER
Sentence-level										
Recurrent neural cascade model [24]	LSTM	IU X-Ray [21]	**79.3**	9.1	0.0	0.0	–	–	–	–
	GRU		78.5	**14.4**	**4.7**	0.0	–	–	–	–
Multi-task learning model [25]		PEIR	0.300	0.218	0.165	0.113	0.149	0.279	–	0.329
Paragraph-level										
Multi-task learning model [25]		IU X-Ray [21]	**0.517**	**0.386**	**0.306**	**0.247**	0.217	**0.447**	–	0.327
Multimodal recurrent model with attention [26]			0.464	0.358	0.270	0.195	**0.274**	0.366	–	–
TieNet [27]			0.2860	0.1597	0.1038	0.0736	0.1076	–	0.2263	–
HRGR-Agent [20]			0.438	0.298	0.208	0.151	–	0.322	–	**0.343**
		CX-CHR	0.673	0.587	0.530	0.486	–	0.612	–	2.895

Table 6
Evaluation metrics (image caption measures).

Metric	Purpose	Algorithm	Strengths	Weaknesses
BLEU [98] 2002	machine translation	N_{gram} *precision*	Correlates with human judgments	Lack of explicit word matching
ROUGE [99] 2004	document summarization	N_{gram} *recall*	Favours long sentences	Works only in single document summarization
METEOR [100] 2005	machine translation	N_{gram} *with synonym matching*	Benefit from synonyms and paraphrase matching	Lack of semantic similarity capturing
CIDEr [101] 2015	image captioning	N_{gram} *with corpu reweighting*	Works in linguistics means	May weight irrelevant sentence's details
SPICE [102] 2016	image captioning	*fobjects * fattributes * frelations*	Can match noun / object between captions	Reliant on the performance of parsing

accuracy, and robustness of the generated annotations.

The original purpose of these common matrices was not to evaluate generated radiology reports. Therefore, some researchers have designed complementary metrices. For instance, a metric called keywords accuracy (KA) calculates accuracy by dividing the number of correctly generated words by the number of ground truth words from the medical text indexer (MTI) annotations [26].

6.2. Qualitative

Qualitative evaluation involves comparing ground truth reports with mode generated reports using content coverage, length, medical term accuracy, and text fluency. For example, [20] utilized Amazon mechanical Turk (MTurk) to conduct surveys. Here, participants chose the generated report that best matched the ground truth report. Jing et al. [25] manually compared the generated paragraphs from their co-attention model with the ground truth to establish which models captured normality and abnormality most efficiently.

7. Discussion and future direction

Deep learning algorithms have the potential to be used in all fields of medicine and could significantly alter the way medicine is practiced. Future DL research should utilize the wealth of medical images and relevant diagnostic reports that are available in PACS to automatically produce clinical reports [13]. Recent attention has focused on generating text reports based on medical data.

Beyond traditional medical image annotation [35,36,51,53,59,80,92] and single sentence-based descriptions [24,25,40], generating radiologist coherent paragraphs has recently attracted researchers [20,25–27]. This presents a more practical and challenging application that can bridge visual medical features with radiologist interpretation. Notably, CNN and RNN have quickly become popular choices for mining radiology images and text, respectively. The main challenge now lies in how to obtain ImageNet-level semantic labels on a large collection of medical images.

Deep learning has several limitations that should be addressed to improve the task of radiology reporting. A reliable reporting system may require tens of millions of image/text samples which are not yet readily available [14]. Furthermore, these samples should be structured without scattered and noisy information to facilitate the learning process for DL models. To date, there are few medical datasets that are large and accessible enough to train multimodal deep CNN. Improving the quantity and quality of radiology data remains an ongoing task.

In a radiology database, the data is unbalanced because abnormal cases are rarer than normal cases. For example, the healthy cases in the IU X-ray chest X-Ray dataset consisted of 2696 images (37%) compared to the 840 images (12%) that represented common diseases and 655 images (9%) that showed less common diseases [24]. Attempted to address this issue by training CNN with different regularization methods including batch normalization and data dropout. In addition, it is challenging to automate labels for medical images as radiologist reports often include ambiguous words. This includes disease prediction rather than if it is present or not [19]. It should be noted that it is difficult to compare various models as researchers conduct their experiments using diverse and sometimes private datasets.

Researchers consider DL as a black box that takes an input, such as a medical image, and generates an output to state a conclusion (e.g. "there is a 0.8 probability of melanoma") without clear explanations [14,105]. This is unacceptable in the medical domain as radiologist need to provide findings as well as underlying justifications. For instance, researchers may attempt to provide the rationale behind the radiologist's description using their proposed models. Considerably more research will need to be conducted to offer reasonable explanations for DL model outcomes.

Most research uses CNN to apply text-image mining in medical imaging. As such, CNN has the widest variety in architecture including AlexNet, VGG-16, GoogLeNet, and ResNet. In the last three years, end-to-end trained CNNs have become the preferred approach for medical imaging interpretation. As such, this could be considered standard practice for mining medical images. In addition, it is likely that the volume of research in leveraging radiology reports for CNN training

will only increase in the near future.

Creating multipurpose reporting systems for radiologists that can detect several diseases simultaneously remains an ongoing challenge. Medical findings often correlate with certain body parts such as the spread of liver metastases and lymph nodes. Despite the promising results of generating radiologist reports, several questions require addressing. For example, what are the clinically related image annotations to be defined? How should the large volume of radiologist images required for DL techniques be labeled? To what extent is the deep CNN framework generalizable for radiology images? Future work should explore valuable semantic diagnostic information and map the many well-written radiologist reports and relevant images.

8. Conclusion

This paper presented a comprehensive literature survey on multimodal datasets to train deep DL models that generate radiology text from images. This field is crucial as these techniques can quickly and accurately provide additional diagnostic criteria by reporting unobservable data from the images and text.

Declaration of Competing Interest

The author declares that they have no conflict of interest.

References

[1] Karpathy A, Fei-Fei L. Deep visual-semantic alignments for generating image descriptions. in Proceedings of the IEEE conference on computer vision and pattern recognition 2015:3128–37.
[2] van Ginneken B, Schaefer-Prokop CM, Prokop M. Computer-aided diagnosis: how to move from the laboratory to the clinic. Radiology 2011;261(3):719–32.
[3] Kohli M, Prevedello LM, Filice RW, Geis JR. Implementing machine learning in radiology practice and research. Am J Roentgenol 2017;208(4):754–60.
[4] Wang J, Yang X, Cai H, Tan W, Jin C, Li L. Discrimination of breast cancer with microcalcifications on mammography by deep learning. Sci Rep 2016;6:27327.
[5] Cheng J-Z, et al. Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans. Sci Rep 2016;6:24454.
[6] McBee MP, et al. Deep learning in radiology. Acad Radiol 2018.
[7] Pouyanfar S, et al. A Survey on Deep Learning: Algorithms, Techniques, and Applications. ACM Computing Surveys (CSUR) 2018;51(5):92.
[8] Alom MZ, et al. The history began from AlexNet. A Comprehensive Survey on Deep Learning Approaches. 2018. arXiv preprint arXiv:1803.01164.
[9] Esteva A, et al. A guide to deep learning in healthcare. Nat Med 2019;25(1):24.
[10] Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE J Biomed Health Inform 2018;22(5):1589–604.
[11] Ravı D, et al. Deep learning for health informatics. IEEE J Biomed Health Inform 2017;21(1):4–21.
[12] Ker J, Wang L, Rao J, Lim T. Deep learning applications in medical image analysis. IEEE Access 2018;6:9375–89.
[13] Litjens G, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88.
[14] Wang F, Casalino LP, Khullar D. Deep learning in medicine—promise, progress, and challenges. JAMA Intern Med 2018.
[15] Akay A, Hess H. Deep learning: current and emerging applications in medicine and technology. IEEE J Biomed Health Inform 2019.
[16] Erickson BJ, Korfiatis P, Kline TL, Akkus Z, Philbrick K, Weston AD. Deep learning in radiology: does one size fit all? J Am Coll Radiol 2018;15(3):521–6.
[17] Lakhani P, et al. Machine learning in radiology: applications beyond image interpretation. J Am Coll Radiol 2018;15(2):350–9.
[18] Imaging and radiology: MedlinePlus Medical Encyclopedia. 2018 [Online]. Available:. https://medlineplus.gov/ency/article/007451.htm.
[19] Shin HC, Lu L, Kim L, Seff A, Yao J, Summers RM. Interleaved text/image deep mining on a large-scale radiology database for automated image interpretation. J Mach Learn Res 2016;17:1–31 Article [Online]. Available:. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84989187487&partnerID=40&md5=83764cf16c1f8dcf723acced65ee2054.
[20] Li CY, Liang X, Hu Z, Xing EP. Hybrid Retrieval-Generation Reinforced Agent for Medical Image Report Generation arXiv preprint arXiv:180508298 2018.
[21] Demner-Fushman D, et al. Preparing a collection of radiology examinations for distribution and retrieval. J Am Med Inform Assoc 2015;23(2):304–10.
[22] Schuyler PL, Hole WT, Tuttle MS, Sherertz DD. The UMLS Metathesaurus: representing different views of biomedical concepts. Bull Med Libr Assoc 1993;81(2):217.
[23] Langlotz CP. RadLex: A New Method for Indexing Online Educational Materials ed Radiological Society of North America; 2006.
[24] Shin H-C, 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 2016:2497–506.
[25] Jing B, Xie P, Xing E. On the Automatic Generation of Medical Imaging Reports arXiv preprint arXiv:171108195 2017.
[26] Xue Y, et al. Multimodal recurrent model with attention for automated radiology report generation. in International Conference on Medical Image Computing and Computer-Assisted Intervention. 2018. p. 457–66.
[27] 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 2018:9049–58.
[28] Thanki RM, Kothari A. Data compression and its application in medical imaging. in Hybrid and Advanced Compression Techniques for Medical Images. 2019. p. 1–15.
[29] Irvin J, et al. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison arXiv preprint arXiv:1901.07031 2019.
[30] Statistics, "Statistics » Diagnostic Imaging Dataset. 2020 [Online]. Available:. https://www.england.nhs.uk/statistics/statistical-work-areas/diagnostic-imaging-dataset/.
[31] Wong HYF, et al. Frequency and distribution of chest radiographic findings in COVID-19 positive patients. Radiology 2020:201160.
[32] Radiology ACo. ACR recommendations for the use of chest radiography and computed tomography (CT) for suspected COVID-19 infection. ACR website. 2020.
[33] Johnson AE, et al. MIMIC-CXR: A Large Publicly Available Database of Labeled Chest Radiographs arXiv preprint arXiv:1901.07042 2019.
[34] 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. in Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on. 2017. p. 3462–71.
[35] Rubin J, Sanghavi D, Zhao C, Lee K, Qadir A, Xu-Wilson M. Large Scale Automated Reading of Frontal and Lateral Chest X-Rays Using Dual Convolutional Neural Networks arXiv preprint arXiv:1804.07839 2018.
[36] Monshi MMA, Poon J, Chung V. Convolutional neural network to detect thorax diseases from multi-view chest X-rays. in Neural Information Processing. iconip-2019. Springer Nature Switzerland AG; 2019. p. 1–11.
[37] Bustos A, Pertusa A, Salinas J-M, de la M. Iglesia-vayá, "PadChest. A Large Chest X-Ray Image Dataset With Multi-Label Annotated Reports. 2019. arXiv preprint arXiv:1901.07441.
[38] Bertrand H, Hashir M, Cohen JP. Do Lateral Views Help Automated Chest X-ray Predictions? arXiv preprint arXiv:1904.08534 2019.
[39] Heath M, Bowyer K, Kopans D, Moore R, Kegelmeyer WP. The digital database for screening mammography. in Proceedings of the 5th International Workshop on Digital Mammography. 2000. p. 212–8.
[40] Kisilev P, Sason E, Barkan E, Hashoul S. Medical image description using multi-task-loss CNN. in Deep Learning and Data Labeling for Medical Applications. Springer; 2016. p. 121–9.
[41] Sahu B, Verma R. DICOM search in medical image archive solution e-sushrut chhavi. 2011 3rd International Conference on Electronics Computer Technology, 6. 2011. p. 256–60.
[42] Six O. The ultimate guide to AI in radiology. Artificial Intelligence in Healthcare Solutions 2019.
[43] Lee SM, et al. Deep learning applications in Chest Radiography and computed tomography: current state of the art. J Thorac Imaging 2019;34(2):75–85.
[44] Mohsen H, El-Dahshan E-SA, El-Horbaty E-SM, Salem A-BM. Classification using deep learning neural networks for brain tumors. Future Comput Inform J 2018;3(1):68–71.
[45] Lam C, Yu C, Huang L, Rubin D. Retinal lesion detection with deep learning using image patches. Invest Ophthalmol Vis Sci 2018;59(1):590–6.
[46] Wang G, et al. Interactive medical image segmentation using deep learning with image-specific fine tuning. IEEE Trans Med Imaging 2018;37(7):1562–73.
[47] Qayyum A, Anwar SM, Awais M, Majid M. Medical image retrieval using deep convolutional neural network. Neurocomputing 2017;266:8–20.
[48] Chaudhari AS, et al. Super-resolution musculoskeletal MRI using deep learning. Magn Reson Med 2018;80(5):2139–54.
[49] Gibson E, et al. NiftyNet: a deep-learning platform for medical imaging. Comput Methods Programs Biomed 2018;158:113–22.
[50] Biswas M, et al. State-of-the-art review on deep learning in medical imaging. Front Biosci (Landmark Ed) 2019;24:392–426.
[51] Shin H-C, Lu L, Kim L, Seff A, Yao J, Summers RM. Interleaved text/image deep mining on a very large-scale radiology database. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2015:1090–9.
[52] Goodfellow I, Bengio Y, Courville A, Bengio Y. Deep Learning. MIT press Cambridge; 2016.
[53] Dong Y, Pan Y, Zhang J, Xu W. Learning to read chest X-ray images from 16000+ examples using CNN. in Proceedings of the Second IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies. 2017. p. 51–7.
[54] Li Y. Deep Reinforcement Learning: An Overview arXiv preprint arXiv:1701.07274 2017.
[55] Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks. in Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics 2011:315–23.
[56] Clevert D-A, Unterthiner T, Hochreiter S. Fast and accurate Deep Network Learning by Exponential Linear Units (elus) arXiv preprint arXiv:1511.07289

2015.
[57] Nwankpa C, Ijomah W, Gachagan A, Marshall S. Activation functions: Comparison of Trends in Practice and Research for Deep Learning arXiv preprint arXiv:1811.03378 2018.
[58] LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc Ieee 1998;86(11):2278–324.
[59] Wang X, et al. Unsupervised Category Discovery Via Looped Deep Pseudo-Task Optimization Using a Large Scale Radiology Image Database arXiv preprint arXiv:1603.07965 2016.
[60] Moeskops P, Viergever MA, Mendrik AM, de Vries LS, Benders MJ, Išgum I. Automatic segmentation of MR brain images with a convolutional neural network. IEEE Trans Med Imaging 2016;35(5):1252–61.
[61] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. in Advances in Neural Information Processing Systems 2012:1097–105.
[62] Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. in European Conference on Computer Vision. 2014. p. 818–33.
[63] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition arXiv preprint arXiv:1409.1556 2014.
[64] Szegedy C, et al. Going deeper with convolutions. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2015:1–9.
[65] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016:770–8.
[66] Xie S, Girshick R, Dollár P, Tu Z, He K. Aggregated residual transformations for deep neural networks. in Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on. 2017. p. 5987–95.
[67] I. ILSVRC2016. http://image-net.org/challenges/LSVRC/2016/results#team (accessed.
[68] Hu J, Shen L, Sun G. Squeeze-and-excitation networks 7. 2017. arXiv preprint arXiv:1709.01507.
[69] Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. Imagenet: a large-scale hierarchical image database. in Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. 2009. p. 248–55.
[70] Stock P, Cisse M. ConvNets and ImageNet beyond accuracy: understanding mistakes and uncovering biases. in Proceedings of the European Conference on Computer Vision (ECCV) 2018:498–512.
[71] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. in Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics 2010:249–56.
[72] Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, inception-resnet and the impact of residual connections on learning. AAAI 2017;4:12.
[73] Lin M, Chen Q, Yan S. Network in network arXiv preprint arXiv:1312.4400 2013.
[74] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017. p. 2261–9.
[75] Williams RJ, Zipser D. A learning algorithm for continually running fully recurrent neural networks. Neural Comput 1989;1(2):270–80.
[76] Mikolov T, Karafiát M, Burget L, Černocký J, Khudanpur S. Recurrent neural network based language model. in Eleventh Annual Conference of the International Speech Communication Association 2010.
[77] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9(8):1735–80.
[78] Cho K, et al. Learning Phrase Representations Using rnn Encoder-Decoder for Statistical Machine Translation arXiv preprint arXiv:1406.1078 2014.
[79] Jia Y, et al. Caffe: convolutional architecture for fast feature embedding. in Proceedings of the 22nd ACM International Conference on Multimedia. 2014. p. 675–8.
[80] Rajpurkar P, et al. Chexnet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning arXiv preprint arXiv:1711.05225 2017.
[81] Abadi M, et al. Tensorflow: a system for large-scale machine learning. in 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16) 2016:265–83.
[82] Paszke A, et al. Automatic Differentiation in Pytorch. 2017.
[83] Deng L, Liu Y. A joint introduction to natural language processing and to deep learning. in Deep Learning in Natural Language Processing. 2018. p. 1–22.
[84] Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. Natural language processing (almost) from scratch. J Mach Learn Res 2011;12(August):2493–537.
[85] Deng L, Liu Y. Deep Learning in Natural Language Processing. Springer; 2018.
[86] He X, Deng L. Deep learning in natural language generation from images. in Deep Learning in Natural Language Processing. 2018. p. 289–307.
[87] Hassanpour S, Langlotz CP. Unsupervised topic modeling in a large free text radiology report repository. J Digit Imaging 2016;29(1):59–62.
[88] Shin H-C, Lu L, Summers RM. Natural language processing for large-scale medical image analysis using deep learning. in Deep Learning for Medical Image Analysis. 2017. p. 405–21.
[89] Zhang Z, Xie Y, Xing F, McGough M, Yang L. Mdnet: a semantically and visually interpretable medical image diagnosis network. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017:6428–36.
[90] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res 2003;3(January):993–1022.
[91] Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016:2921–9.
[92] Wang H, Xia Y. Chestnet: A Deep Neural Network for Classification of Thoracic Diseases on Chest Radiography arXiv preprint arXiv:1807.03058 2018.
[93] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: visual explanations from deep networks via gradient-based localization. in Proceedings of the IEEE International Conference on Computer Vision 2017:618–26.
[94] Howard J, Ruder S. Universal Language Model Fine-Tuning for Text Classification arXiv preprint arXiv:1801.06146 2018.
[95] Peng Y, Wang X, Lu L, Bagheri M, Summers R, Lu Z. NegBio: a high-performance tool for negation and uncertainty detection in radiology reports. AMIA Summits on Translational Science Proceedings 2018;2017:188.
[96] Xu K, et al. Show, attend and tell: neural image caption generation with visual attention. in International Conference on Machine Learning 2015:2048–57.
[97] Guo J, Yuan X, Zheng X, Xu P, Xiao Y, Liu B. Diagnosis labeling with disease-specific characteristics mining. Artif Intell Med 2018;90:25–33.
[98] Papineni K, Roukos S, Ward T, Zhu W-J. BLEU: a method for automatic evaluation of machine translation. in Proceedings of the *40th Annual Meeting on Association for Computational Linguistics*. 2002. p. 311–8.
[99] Lin C-Y. Rouge: a package for automatic evaluation of summaries. Text Summarization Branches Out. 2004.
[100] Denkowski M, Lavie A. Meteor universal: language specific translation evaluation for any target language. in Proceedings of the Ninth Workshop on Statistical Machine Translation 2014:376–80.
[101] Vedantam R, Lawrence Zitnick C, Parikh D. Cider: consensus-based image description evaluation. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2015:4566–75.
[102] Anderson P, Fernando B, Johnson M, Gould S. Spice: semantic propositional image caption evaluation. in European Conference on Computer Vision. 2016. p. 382–98.
[103] Lin C-Y, Och FJ. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. in Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. 2004. p. 605.
[104] Kilickaya M, Erdem A, Ikizler-Cinbis N, Erdem E. Re-evaluating automatic metrics for image captioning arXiv preprint arXiv:1612.07600 2016.
[105] Hicks SA, et al. Comprehensible reasoning and automated reporting of medical examinations based on deep learning analysis. in Proceedings of the 9th ACM Multimedia Systems Conference. 2018. p. 490–3.