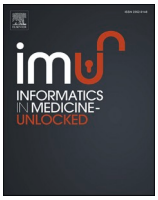
Informatics in Medicine Unlocked 39 (2023) 101273

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

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



Deep learning approaches to automatic radiology report generation: A systematic review

Yuxiang Liao, Hantao Liu, Irena Spasić *

School of Computer Science and Informatics, Cardiff University, UK

ARTICLE INFO

Keywords:
Image processing
Natural language generation
Natural language processing
Deep learning
Neural network

ABSTRACT

Background: A radiology report communicates the imaging findings to the referring clinicians. The rising number of referrals has created a bottleneck in healthcare. Writing a report takes disproportionally more time than the imaging itself. Therefore, Automatic Radiology Report Generation (ARRG) has a great potential to unclog this bottleneck.

Objectives: This study aims to provide a systematic review of Deep Learning (DL) approaches to ARRG. Specifically, it aims to answer the following research questions. What data have been used to train and evaluate DL approaches to ARRG? How are DL approaches to ARRG evaluated? How is DL used to generate the reports from radiology images?

Materials and methods: We followed the PRISMA guidelines. We retrieved 1443 records from PubMed and Web of Science on November 3, 2021. Relevant studies were categorized and compared from multiple perspectives. The corresponding findings were reported narratively.

Results: A total of 41 studies were included. We identified 14 radiology datasets. In terms of evaluation, we identified four commonly used natural language generation metrics, six clinical efficacy metrics, and other qualitative methods. We compared DL approaches with respect to the underlying neural network architecture, the method of text generation, problem representation, training strategy, interpretability, and intermediate processing.

Discussion and conclusion: Data imbalance (normal versus abnormal cases) and the inner complexity of reports pose major difficulties in ARRG. More appropriate evaluation metrics are required as well as datasets on a much larger scale. Leveraging structured representation of radiology reports and pre-trained language models warrant further research.

1. Introduction

1.1. Background

Radiology tests based on modalities such as X-ray, ultrasound (US), computed tomography (CT), and magnetic resonance imaging (MRI) provide detailed insights into the patients' bodies without them needing to undergo invasive explorative surgeries. They can help screening for and diagnosis of medical conditions as well as monitoring the response to treatments. As such, radiology tests remain the most common types of imaging tests. As many as 45.2 million imaging tests were reported in England between September 2018 and September 2019, the top four tests were the above radiology tests, which accounted for 96% of the imaging tests [1]. Before the pandemic in September 2019, more than 9000 people were waiting for CT and MRI test for at least six weeks. In just a year, this number became almost ten times higher [2]. The volume of imaging referrals is continually increasing to the point that many departments promote the idea of 30–60 min turn-around times in order to stay competitive [3]. However, there were only 12.8 radiologists per million population in Europe in 2020, and the corresponding number in the UK was even smaller [2]. The issues of enormous daily diagnostic needs and the lack of radiologists are further aggravated by problems such as diagnostic errors [4–6] and interpretation discrepancies between radiologists and physicians [7].

An imaging report represents the most important means of communication between a radiologist and the referring medical professional, both serving an effort to provide a high-quality patient care [8]. Some of the most important skills any radiologist needs are observational skills to

* Corresponding author.
E-mail addresses: liaoy11@cardiff.ac.uk (Y. Liao), liuh35@cardiff.ac.uk (H. Liu), spasici@cardiff.ac.uk (I. Spasić).

https://doi.org/10.1016/j.imu.2023.101273
Received 20 December 2022; Received in revised form 10 May 2023; Accepted 12 May 2023
Available online 19 May 2023

identify abnormalities, analytical skills to relate observed abnormalities to the underlying pathology and communicative skills to convey their interpretation clearly to both clinicians and patients [8]. A widespread shortage of these skills [9] emphasizes the need for automation in this area, which leads us to the task of Automatic Radiology Report Generation (ARRG).

1.2. The goal of ARRG

ARRG is a specific application of Automatic Image Captioning (AIC) in the medical radiology domain. As a form of image-to-sequence generation, AIC relies on understanding images such as scenes, objects, object properties, and their interactions; and producing textual descriptions that are both syntactically and semantically sound [10]. ARRG focuses solely on radiology images, with an emphasis on recognizing normal and abnormal appearances and describing them accurately and comprehensively. According to Kaur et al. [7], a high-quality automated radiology report should 1) be clear, concise, and structured for the referring clinician to read effortlessly; 2) be complete regarding positive and negative observations; 3) prioritize the observations; 4) use uniform language (medical terminology) to prevent ambiguity; 5) follow the conventional report format used by radiologists.

To achieve this, attempts in this field involve generating narrative reports [11–14] and structured reports [15–17], and retrieving appropriate sentences from pre-constructed retrieval databases [18–21], as shown in Figs. 1 and 2. We define narrative radiology reports as direct descriptions of all relevant information and diagnostic impressions provided by radiologists with free or structured layouts. It is the most common report type in publicly available radiology image/text datasets. The narrative reports vary excessively in language, length and style, which may affect their clarity and hence the referring clinicians' decision-making [22]. These issues gave rise to the idea of structured reporting, which has the potential to improve the clarity of radiology reports. Structured radiology reports use uniform, standardized, organized terms to describe the medical content without being affected by the reporting style of radiologists [23]. These terms are considered as structured report entities that can be easily converted into natural-sounding sentences via templates [24], such as a set of tuples including anatomy, anatomy qualifier, observation, observation qualifier, certainty and negation [25,26] and common data elements for radiology [27,28]. However, the existing surveys of ARRG have either reviewed only a few studies [29–31] or focused on a narrower scope [7], and none of them took into consideration the structured report generation, leaving room for a more comprehensive review of this area.

2. Deep learning approaches to ARRG

One of the earliest studies towards ARRG dates back to 2015 when Shin et al. [32] trained a deep Convolutional Neural Network (CNN) to generate keywords based on CT/MRI images. Later, Shin et al. [33] went on to design the first ARRG system, which could generate five keywords from chest X-rays concerning disease location, severity, and the affected anatomical sites. In 2018, research on ARRG systems started to gain widespread attention [12,14,20,34]. Further details about the evolution of relevant techniques are provided in Fig. 1.

The ARRG models typically rely on Deep Learning (DL) approaches, which have shown promising results in AIC [10]. DL techniques enable the models to capture complex patterns and relationships directly from raw data. In contrast, non-DL approaches, such as conventional machine learning, necessitate manually engineered features to develop mathematical or statistical models for pattern recognition [35]. The design and extraction of such features require domain knowledge [36], resulting in the formation of multi-disciplinary knowledge barriers. Today, non-DL approaches are rarely used as standalone methods in ARRG [37]. Instead, they are often employed in combination with DL methods by which the features are extracted to avoid the time-consuming manual feature engineering [18–21].

On the other hand, there is increasing literature showing that DL plays an important role in many single-modality tasks in healthcare, such as understanding and interpreting health records [38,39], and handling various medical imaging tasks (e.g. image segmentation) [40]. However, as a multi-modality generative task that involves computer vision [41], Natural Language Processing (NLP) [42] and medical image analysis [43], ARRG is a computationally challenging problem. To comprehensively discuss the DL approaches in ARRG, this review is conducted from three key aspects, including training datasets, model designs and evaluation methods, which correspond to the targets, approaches, and outcomes of the model training, respectively.

3. Materials and Methods

This review was conducted in accordance with the Preferred Reporting Items for Systematic Review and Meta-Analyses (PRISMA) 2020 statement [44]. The main aim of this study is to systematically review DL approaches to ARRG in order to answer the following Research Questions (RQs).

RQ1: What data have been used to train and evaluate DL approaches to ARRG?

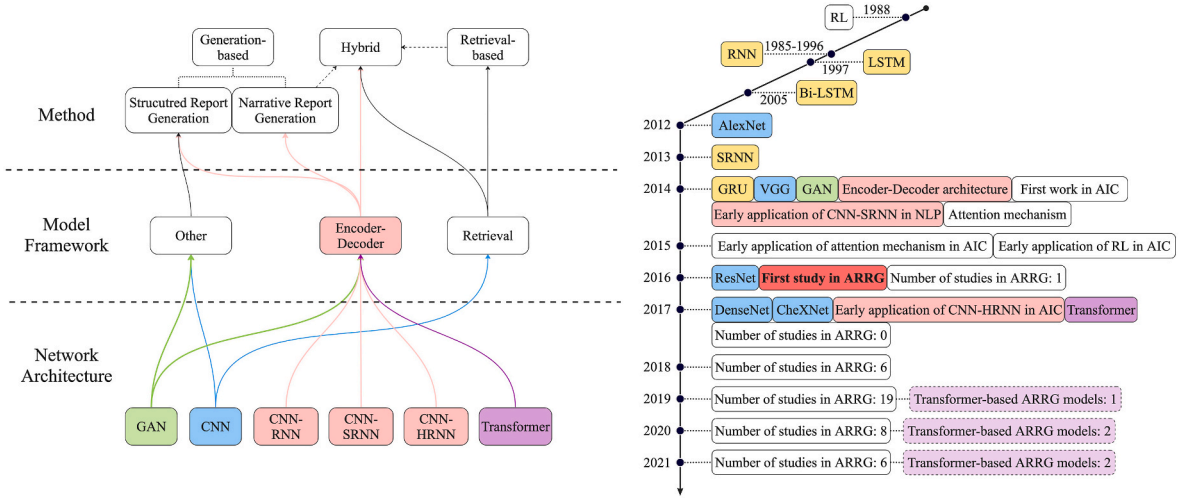


Fig. 1. A diagram showing the hieratical relationships between ARRG methods, framework, and architectures (left) and a timeline in terms of the milestones of the technologies involved in ARRG (right).

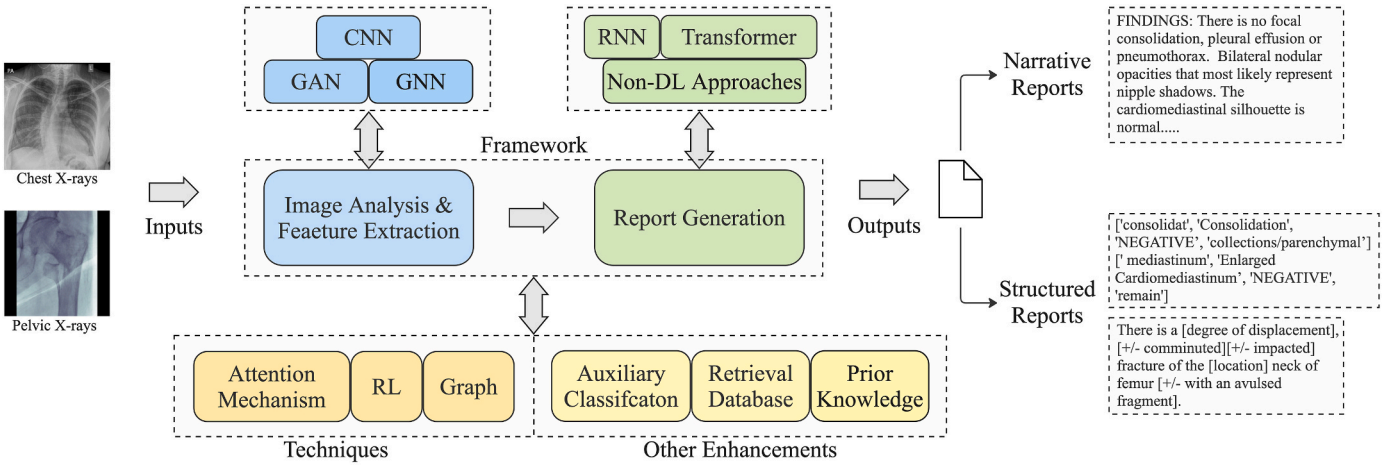


Fig. 2. Overview of the ARRG system workflow.

RQ2: How is DL used to generate the reports from radiology images?
RQ3: How are DL approaches to ARRG evaluated?

RQ1 aims to identify the key properties of data used to train and evaluate DL approaches to ARRG. These properties include imaging modalities used, anatomical sites involved, the internal structure of the reports, their labels and basic statistical properties. RQ2 describes the methods used to solve the problem of ARRG. This question is concerned with the way in which ARRG can be represented formally by mapping it onto a set of relevant computational problems (e.g. object detection, multi-label classification, text generation) so that any future research on ARRG can consider relevant literature that can be useful for but is not strictly limited to the ARRG. More importantly, it addresses the way in which they are integrated into a DL framework to support ARRG. Finally, RQ3 focuses on the evaluation methods for ARRG. The multi-modal nature of this problem makes it difficult to compare the effectiveness of different approaches, thus highlighting the need for a comprehensive evaluation framework.

The scope of the review is defined by a set of inclusion and exclusion criteria described in Table 1. Relevant studies were identified from two scientific databases, Web of Science [45], which comprises 171 million citations in various academic disciplines, and PubMed [46], which indexes more than 33 million citations on the subject of biomedical and life sciences. The process of constructing the search query is shown in Table 2. The query is built on top of three facets: “deep learning”, “text generation”, and “radiology”. They correspond to the method, target, and application area of ARRG, respectively. The three facets were combined into a Boolean search query using the AND operator.

Table 1
Inclusion and exclusion criteria.

No.	Inclusion Criteria
1	Studies that use DL to generate radiology reports or labels (structured report entities) but only if the labels contain sufficient information allowing them to be easily expanded into a full report.
2	Studies that apply data fusion of radiology images and radiology reports as part of training.
No.	**Exclusion Criteria**
1	The text of the radiology report (either input or output) is written in a language other than English.
2	Studies that are not original research, e.g. a review.
3	Studies that have not undergone scrutiny procedures, e.g. peer review.

Table 2
Search queries for PubMed and Web of Science.

Query	PubMed	Web of Science
1	("deep learning"[All Fields] OR "neural networks"[All Fields])	(deep learning OR neural networks)
2	("natural language processing"[All Fields] OR NLP[All Fields] OR "natural language generation"[All Fields] OR (("report*" OR "finding*" OR "text") AND "generat*") OR ("image*" AND ("reporting" OR "captioning")))	(natural language processing OR natural language generation OR ((report$ OR finding$ OR text) NEAR/10 (generat*)) OR (image$ NEAR/10 (reporting OR captioning)))
3	(radiology[All Fields] OR radiography[All Fields] OR "computed tomography"[All Fields] OR CT[All Fields] OR "positron emission tomography"[All Fields] OR PET[All Fields] OR "magnetic resonance imaging"[All Fields] OR MRI[All Fields] OR X-ray[All Fields] OR ultrasound[All Fields] OR fluoroscopy[All Fields] OR mammography[All Fields] OR "nuclear medicine"[All Fields])	(radiology OR radiography OR computed tomography OR CT OR positron emission tomography OR PET OR magnetic resonance imaging OR MRI OR X-ray OR ultrasound OR fluoroscopy OR mammography OR nuclear medicine)

4. Results

4.1. Study selection

The search conducted on November 3, 2021, returned a total of 1443 records. All titles and abstracts were screened by two independent reviewers (Y.L., I.S.), achieving high interrater agreement measured by Cohen’s kappa coefficient ($K = 0.844$, $n = 1443$) [47]. All disagreements were resolved by the third independent reviewer (H.L.). One reviewer (Y.L.) reviewed the full text of the 34 candidate studies. The inspection of their references revealed 21 additional studies, which were added to the pending list for full-text reviewing. Any uncertainties were resolved by discussion (Y.L., I.S., H.L.). Ultimately, a total of 41 studies were included. The selection process and its outcomes are summarized in Fig. 3. Data were extracted by one reviewer (Y.L.). After deeming meta-analysis not applicable to this review due to the heterogeneity of training data and evaluation methods, we conducted a narrative synthesis of the main findings.

4.2. RQ1: Data

The quality of training data is an important factor affecting the performance of DL models. Table 3 summarizes publicly available

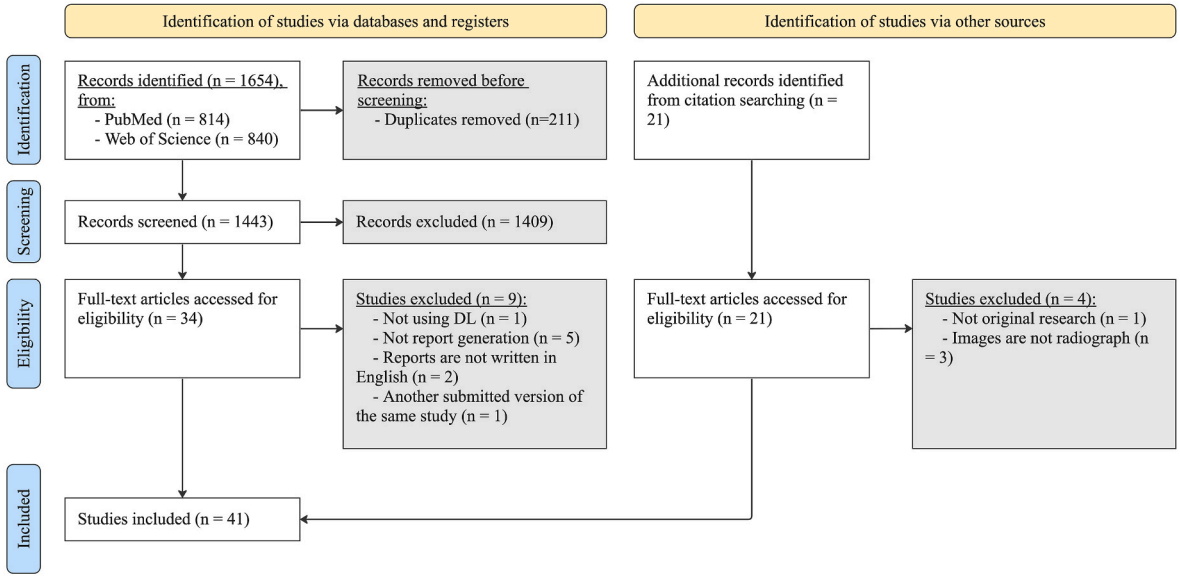


Fig. 3. PRISMA flow diagram of search strategy and study selection.

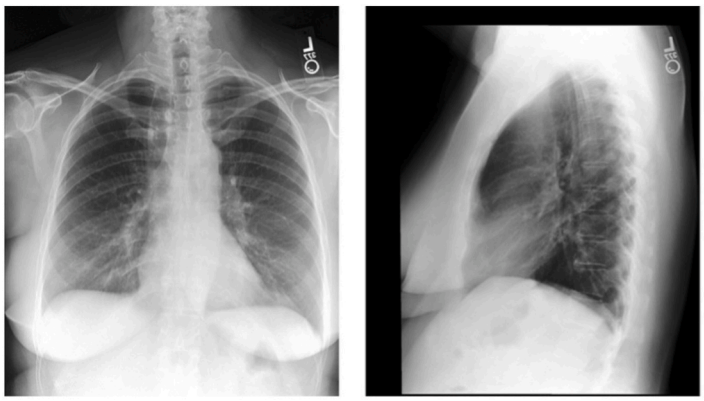
Table 3
A summary of publicly available radiology image/text datasets used in training the ARRG systems.

Dataset	Image				Text				Used by
	Modality	Size	Format	Type	Size	Type	Report Format	Description	
IU X-ray (2016) [49]	Chest X-ray	7470	PNG, DICOM	Frontal and lateral views	3955	Semi-structured reports + MeSH tags[a]	XML	The reports mainly contain four headings: comparison, indication, findings, impression.	[11–14,19–21,33,34,50–65]
MIMIC-CXR (2019) [66,67]	Chest X-ray	377,110	JPEG, DICOM	Frontal and lateral views	227,835	Semi-structured reports + Labels	TXT	The reports mainly contain seven sections as in Fig. 5. The labels are automatically mined.	[13,16,18,19,59,63,68,69], [57] (generalisation measurement)
INbreast (2012) [70]	Mammography	410	DICOM	CC and MLO views[b] with contours	117	Free-text reports + Labels	TXT	The labels consist of annotations and classification labels of lesions.	[71]
DeepLesion (2018) [72]	CT	32,120	PNG	CT slides with 32,735 lesion bounding boxes	22,842	Structured reports	Not available	The reports consist of 171 unique labels cover organs, lesions, and the corresponding shape, size, location.	[15,73]
Liver CT Annotation (2015) [74]	Liver CT	50	MAT	3D CT images with liver masks and bounding boxes	50	Structured reports	RDF	The reports consist of 73 classification questions with either a close-ended answer or an open-ended numerical/ narrative answer.	[75]
PEIR Gross (2018) [12]	21 categories	7442	JPEG	Pathology images	7442	Free-text reports	Plain text	Short captions.	[12]

Note: If not specified, the annotations of images and text (e.g. bounding boxes, labels and reports) were manually generated by radiologists.
[a] MeSH [76] is a biomedical and health-related thesaurus, in which the terms are hierarchically organized and have synonyms.
[b] Craniocaudal (CC) view and mediolateral oblique (MLO) view.

datasets used to train the ARRG systems. The imaging data were generated using four prevalent modalities, including X-rays, CT, MRI, and US. Some of the report have no specific structure, some are semi-structured by sections by means of headings, while others are fully structured as tuples, which can be extended into full reports by either rule-based or DL approaches [48]. Fig. 4 provides an example of a chest X-ray study, where the findings and impression are typically the primary targets for ARRG. The *findings* section describes a radiologist's observations regarding different regions in the image, whereas the *impression* section summarizes these observations. A comprehensive list of headings is provided in Fig. 5.

Some of the datasets used in the reviewed studies were not publicly available or were found not to be appropriate for the ARRG tasks when used on their own. For example, **ChestX-ray8/ChestX-ray14** [77] and **CheXpert** [78] contain only diseases labels rather than reports, which were typically used to pre-train [20,21,51,54,56,62,65,68] or fine-tune [61] a model, or used in combination with other datasets [34,58]. Other datasets used in the included studies were not publicly available, including MRI datasets [17,79,80], US datasets [11,81–83], a chest X-ray dataset [84], and a mammography dataset [71].



EXAMINATION: CHEST (PA AND LAT)
INDICATION: History:
breath
F with shortness of
TECHNIQUE: Chest PA and lateral
COMPARISON:
FINDINGS: The cardiac, mediastinal and hilar contours are normal. Pulmonary vasculature is normal. Lungs are clear. No pleural effusion or pneumothorax is present. Multiple clips are again seen projecting over the left breast.
Remote left-sided rib fractures are also re-demonstrated.
IMPRESSION: No acute cardiopulmonary abnormality.

Fig. 4. An example of radiology images with the corresponding report from the MIMIC-CXR dataset [66].

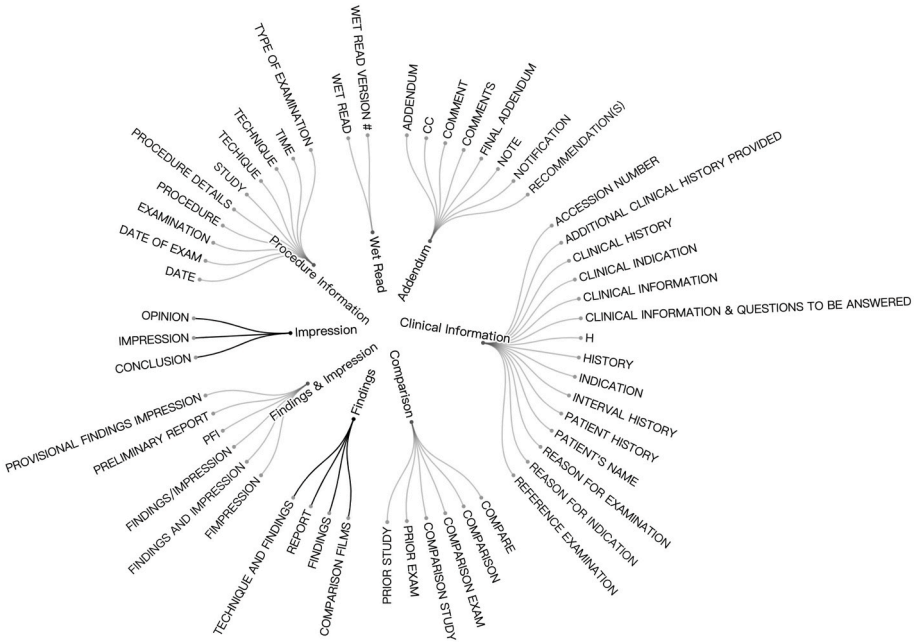


Fig. 5. The common sections (the inner circle) and a few corresponding headings (the outer circle) in semi-structured reports. Extracted from the MIMIC-CXR dataset [66].

4.3. RQ2: Deep learning approaches

4.3.1. Overview of frameworks, network architectures, and techniques

ARRG is a branch of AIC applied in radiology. The majority of ARRG models were derived from one or more AIC models and were further enhanced to meet specific application requirements. This section gives a broad overview of the fundamental frameworks, architectures, and techniques involved in the ARRG models to help researchers build a general picture regarding ARRG. We refer the reader to the recent AIC review study for the details [10].

4.3.1.1. Frameworks for developing ARRG models

4.3.1.1.1. Encoder-decoder framework. The encoder-decoder architecture is a basic framework for developing DL-based, end-to-end AIC models [10]. It was originally introduced for sequence-to-sequence generation [85] and later adopted for image-to-sequence generation in the AIC field [86]. This framework consists of two core components: a visual encoder that extracts image features and a textual decoder that learns the mapping from the image representation to text representation and consequently generates sequences. In ARRG, the generated sequences may consist of narrative words [11–14,20,21,34,50–52,54–65,69,71,81–83] or structured report entities [15–17,33,53,68,73,75,79,80,87,88]. Moreover, the encoder and decoder can be implemented by different network architectures, as shown in Fig. 1. For example, CNN-RNN architecture [11,33,34,53,65,68,71,81–83] is one of the implementations, of which CNN serves as the encoder and RNN serves as a decoder.

4.3.1.1.2. Retrieval framework. A less commonly seen framework in ARRG is the retrieval framework, which has no fixed architecture [18–21]. The key concern of using this framework is the design of retrieval methods that match the extracted image feature to corresponding sentence templates, which raises two other considerations – the methods for visual feature extraction and the construction of template databases for retrieval. We identified four retrieval methods in ARRG, including computing cosine similarity between the visual embeddings of images and choosing the corresponding sentences [19]; aligning the visual and semantic features and computing the visual-semantic similarities via an attention-weighted sum of squared l2-normalized Euclidean distance [18]; treating sentence selection as a multi-label classification problem [21]; or training an agent to retrieve sentence via reinforcement learning [20].

4.3.1.1.3. Other frameworks. Some researchers have approached the

ARRG problem by transforming it into a multi-label classification problem, using various frameworks [15–17,73,75,79,80,87]. These frameworks can be divided into two broad categories, one of which is utilizing CNN-based methods for fine-grained label classification [15,16,73,75,80,87], while the other leverages Generative Adversarial Network (GAN)-based methods for classification accompanied by image segmentation [17,79]. The generated labels are pre-designed structured report entities that can be easily compiled into narrative reports via different approaches, such as using pattern matching [16], decision trees [80,87], and symbolic logic reasoning [17].

4.3.1.2. Network architectures used in ARRG model frameworks

4.3.1.2.1. Convolutional Neural Network (CNN). CNNs are feed-forward neural networks with convolution layers that operate on adjacent pixels to extract features from images, such as edges, shapes, and textures. In ARRG, CNNs are widely used for visual feature extraction in various frameworks. Fig. 6 illustrates the extraction processes, which produce two categories of visual features. The first category is global image features that are typically extracted using pre-trained CNNs. These features could be in the form of a global feature vector from the last pooling layer [11,16,19,20,33,50,53,55,57,60,68,71,75,82,83,87] or a matrix of spatial image features reshaped from a set of feature maps [12,13,17,18,21,34,52,56,58,61–64,69,88], or a combination of both [14,54,59]. The spatial feature matrix facilitates the attention mechanism to better attend to various spatial locations [64]. The second category is regional image features of areas of interest detected by CNN detectors [11,15,51,65,73,81]. Apart from serving as visual encoders, CNNs can also be used for labelling images, with generated labels either used to promote the text generation module to conduct longer reports [12,34,50,54,56,60,64,65,71] or compiled into reports if they correspond to informative structured report entities [15,16,73,75,80,87].

4.3.1.2.2. Recurrent neural network (RNN). RNN and their variants (e.g. LSTM [89] and GRU [90]) can maintain long-range sequential information in their hidden state, ensuring that each word is generated according to its context. In ARRG, they are typically integrated with CNNs and responsible for textual decoding, forming the basic CNN-RNN architecture [11,33,34,53,65,68,71,81–83]. Moreover, there are two modified branches of CNN-RNN architecture, including CNN-SRNN [20,55,88] and CNN-HRNN [12,14,52,54,56–61,64]. CNN-SRNN employs stacked RNNs (SRNN) as the decoder. Compared with general RNNs, SRNN has multiple recurrent hidden layers stacked on top of each other, increasing the observation and capture of sequence inputs at different time scales, thus allowing a more natural representation of sequence text [91,92]. The SRNN structure is also feasible to be integrated into more complex CNN-HRNN architecture [14,57,58]. On the other side, CNN-HRNN uses hierarchical RNN (HRNN) as the decoder. HRNN stacks multiple RNNs in a way that models the hierarchical structure of text sequence, enabling it better capturing linguistic features and generate longer texts [93]. Generally, text is hierarchically structured into sentences and words. Therefore, the decoding process typically starts with a sentence decoder that generates sentence-level semantic features (also known as topic vectors), followed by a word decoder that parses the topic vector into a sequence of words [12,52,54,60,64]. In addition, the radiology report structure also gives rise to a hierarchy that HRNN can leverage [14,57,58].

4.3.1.2.3. Transformer. The transformer architecture is based on the encoder-decoder framework and exclusively employs self-attention units [94]. This design makes the training parallelizable, leading to a greater computational efficiency, and enable better capture of long-range dependencies in sequences. Although the transformer encoder and decoder can be split and composited with other architectures, such as CNN encoder or RNN decoder [10], the entire transformer architecture is usually connected to a CNN encoder in ARRG [13,62,63,69]. More details about the self-attention unit are discussed in Section 3.3.1.3.3.

4.3.1.2.4. Generative Adversarial Network (GAN). GAN [95] aims to train a generator network that can generate new data resembling a given training dataset. This is accomplished through an adversarial training process, where an additional discriminator network is introduced to work against the generator network. Specifically, the generator network

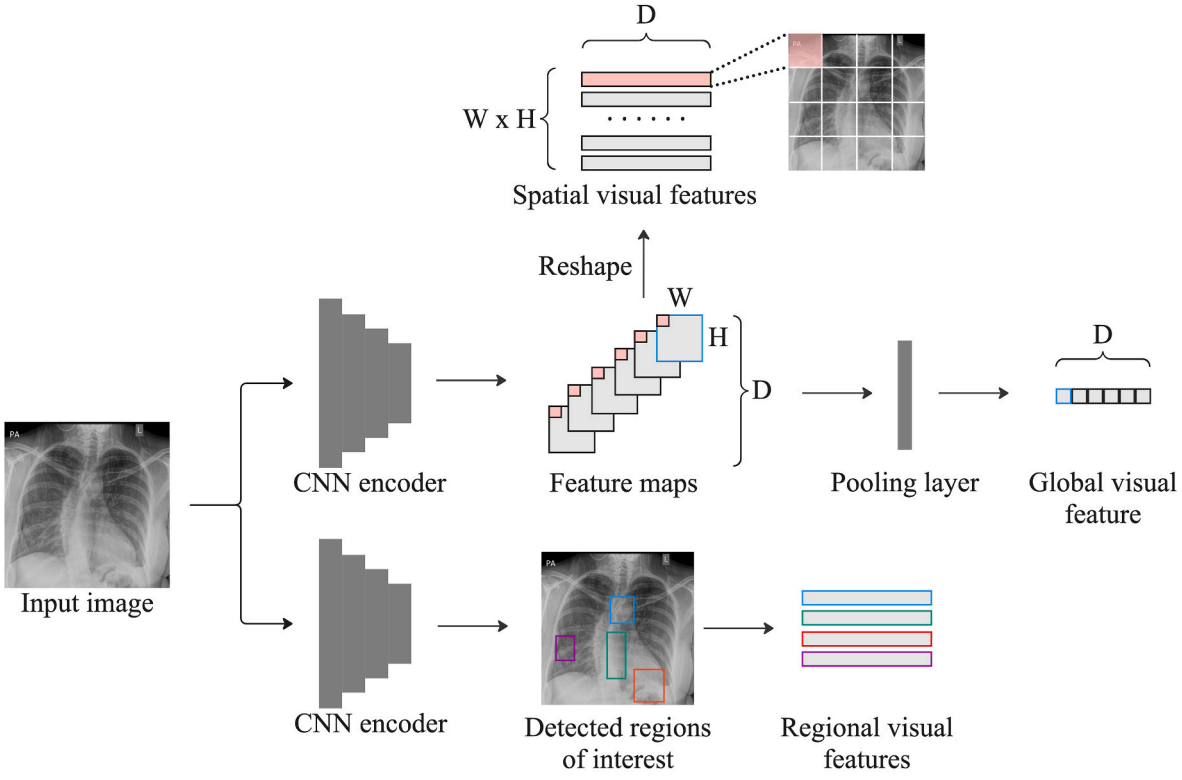


Fig. 6. The visual extraction processes of using CNNs. The upper two paths indicate the global feature encoding. The lower path shows the regional feature encoding.

is trained to generate realistic data that can evade detection by the discriminator network, while the discriminator network is trained to distinguish between the generated data and the real data. In ARRG, GAN is typically employed for the segmentation of the spinal structures in lumbar spine MRI [17,79]. Nevertheless, one study innovatively used the inverse mapping of the GAN's generator instead of the traditional CNN encoder for visual extraction [50].

4.3.1.3. Techniques to enhance ARRG models

4.3.1.3.1. Graph structure. The graph is a useful structure for explicitly representing the relationships between entities, which consists of nodes and edges. Both image and text can be encoded as a graph [21]. For example, an image can be converted into a graph by treating pixels as nodes and linking adjacent pixel with edges. Similarly, in text, individual words can be assigned as nodes and the relationships among words can be represented by edges. In ARRG, Graph Neural Networks (GNNs) such as the Graph Convolutional Network (GCN) and Graph Transformer (GTR) have been utilized to leverage the graph structure [21,56]. Fig. 7 shows a simplified framework that incorporates GNNs for learning graph representations. In ARRG, graph structures are typically designed based on prior knowledge of radiology ontology and constructed from corresponding reports [21,56]. Conversely, in AIC, graphs are constructed using object detection and relationship prediction [10, 96] or directly by off-the-shelf scene graph parsers [97,98]. Graph structures have also been used to improve the consistency of spinal structure classification [17]. In this case, prior knowledge of spinal structures was converted into a graph and embedded into the model to enable reasoning capabilities.

4.3.1.3.2. Reinforcement learning (RL). RL is a machine learning paradigm that aims to train an agent to interact with an environment with optimal actions [99]. In supervised learning, ARRG is commonly performed by minimizing the cross-entropy loss via gradient descent, allowing the model to fit the data. However, this approach may not necessarily allow the model to optimize toward a specific metric of interest. In contrast, RL can directly use metrics as rewards and optimize the model by policy gradient, alleviating the discrepancy between the model training goal and a given evaluation metric. In ARRG, RL has been combined with different architectures, including the CNN-HRNN architecture [20,59,61] and CNN-transformer architecture [51], in which the REINFORCE algorithm [100] is the most commonly used policy-gradient method. When RL is introduced, the ARRG problem is redefined as follows: the *agent* refers to an ARRG model; the *environment* is the input of the model (i.e. the visual features and the input sequences); the *policy* is the model's parameter; the model's outputs indicate a sequence of *actions* taken by the agent under the current policy; the ground-truth reports define the optimal sequences of actions; and the *rewards* are obtained by comparing the agent's actions and the optimal actions via the target metrics.

4.3.1.3.3. Attention mechanism. The attention mechanism is a method that can combine the elements of distinct feature embeddings with different weights according to element-wise correlation rather than relying solely on a fixed representation [101,102]. In ARRG, this method can be divided into cross-model attention (CMA) [12,14,17,18,20,34, 52,54,56–61,64,65,68,69,88] and intra-model attention (IMA) [13,21, 34,51,62,63,69]. The major difference is that the feature embeddings in CMA come from distinct models, whereas those in IMA are the same embedding from a single model. Furthermore, CMA typically equips the soft attention mechanism [101,102] to establish dynamic associations between visual features and linguistic features [14,18,34,52,54,56–59, 61,64,65,68,69,88]. Several soft-attention improvements have been proposed in ARRG to capture more information from various aspects, such as local semantic attention, which discloses the dependency of local visual features and distinct symbolic nodes [17]; multi-attention, which refines the visual attention process into channel- and position-level processing [52]; and co-attention, which enables the linguistic features to simultaneously attend to visual features and predicted label features [12,60]. On the other hand, IMA is typically combined with CMA, following the transformer architecture [13,21,51,62,63,69]. In transformer, both types of attention are based on the scaled dot-product attention mechanism (also known as self-attention) [94]; IMA captures the internal dependencies of the feature embeddings in the encoder and the decode, and the resulting weighted feature representations are correlated by the processing of CMA. In addition to the sequential usage in the transformer architecture, IMA can also be used in parallel with CMA, of which both weighted features are concatenated and used for auxiliary classification [34]. Fig. 8 illustrates the common usages of

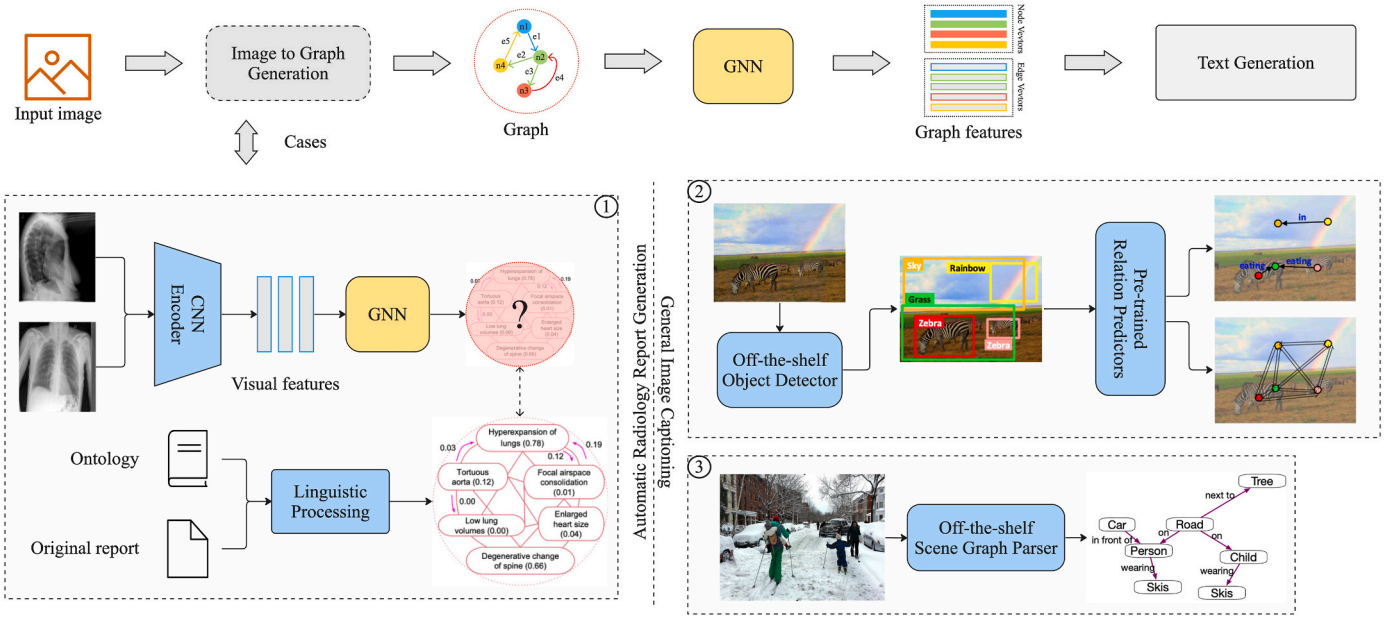


Fig. 7. Overview of encoder-decoder framework integrated with GNN for graph embedding. The construction processes of graphs are demonstrated in three cases. Case 1 is training a GNN to learn to generate graphs for ARRG [21]. Cases 2 and 3 are found in general image captioning. Case 2 uses pre-defined predictors to extract the semantic and spatial relationships from the detected object and forms them in graph structures [96]. Case 3 uses an off-the-shelf graph parser to generate scene graphs [97].

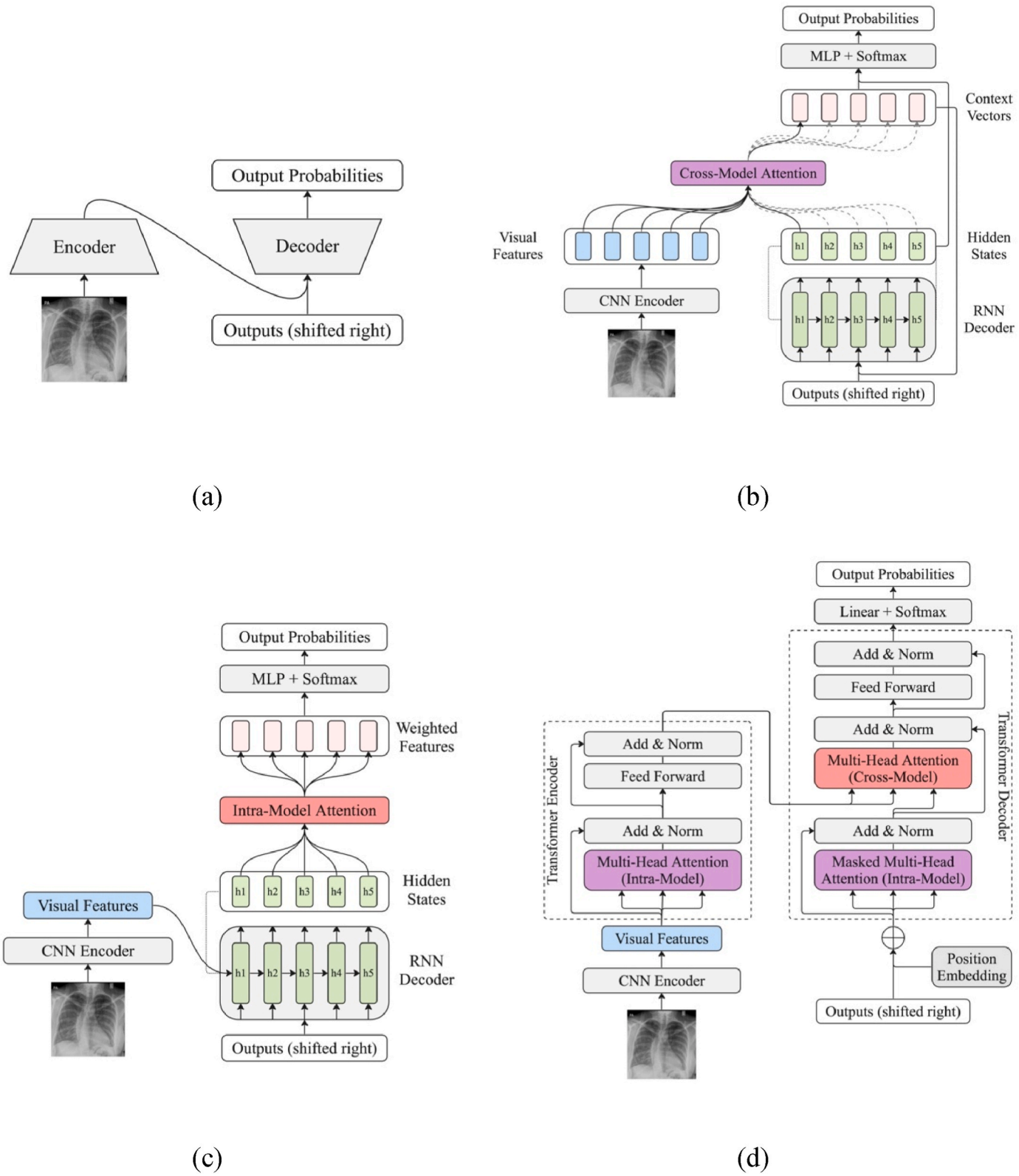


Fig. 8. The general structures of ARRG models that utilize attention mechanisms. We use the encoder-decoder framework as examples: (a) without attention mechanism; (b) with cross-model attention; (c) with intra-model attention; (d) with both cross- and intra-model attention n.

CMA and IMA in ARRG.

4.3.2. *Targeting the report generation*

Existing studies have been proposed to address the ARRG problem by transforming it into specific DL tasks that cater to different objectives and requirements based on the application scenarios (training datasets). These approaches have resulted in the development of various ARRG models, which can be broadly classified into three categories, as depicted in Fig. 9. The most salient features of these ARRG models are illustrated in Appendix A.

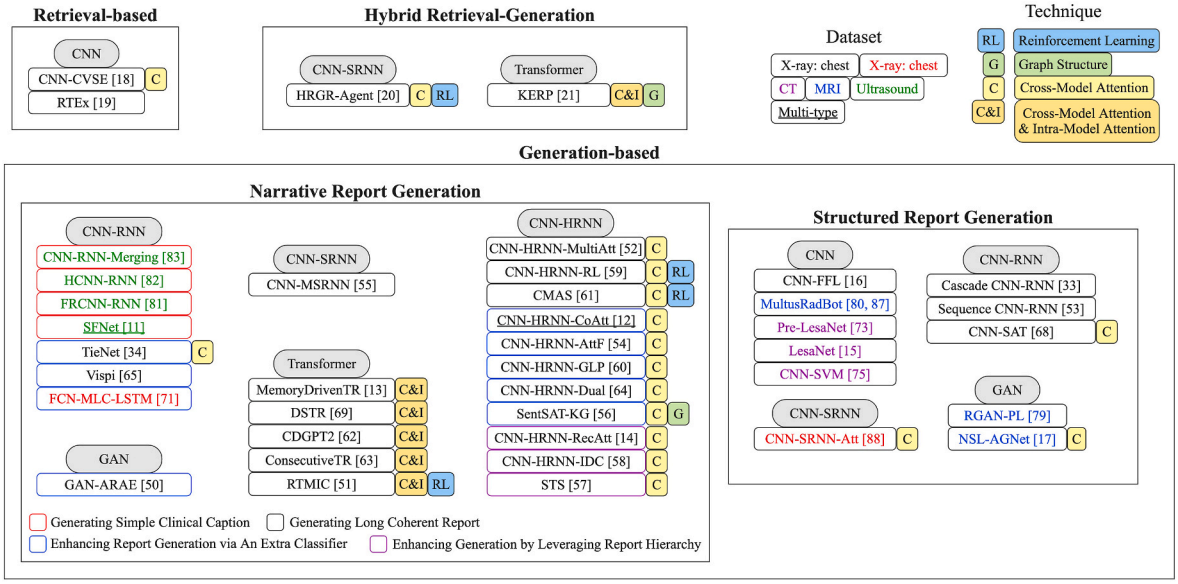


Fig. 9. Fine-grained classification of ARRG systems. The systems are further distinguished by colour and symbols according to different features. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

4.3.2.1. *Narrative report generation*

4.3.2.1.1. Simple clinical caption. Narrative report generation is the most prevalent objective of ARRG models. However, the expected forms of the generated reports would affect the choice of the model architecture. In the case of generating US reports, which tend to be short descriptions or voice-over captions explaining the image, the corresponding ARRG models are designed using a simple CNN-RNN architecture [11,81–83]. CNN-RNN-Merging [83] simply concatenated the feature vectors of the CNN encoder and RNN decoder and passed them to a fully connected layer to predict the following words. HCNN-RNN [82] proposed an ensemble of multiple CNNs to cope with their multi-class dataset. FRCNN-RNN [81] and SFNet [11] captured the location and semantic information of focus areas, producing overall representations that were subsequently concatenated with text features for report generation. Notably, SFNet fused the features of focus areas at a different time node, achieving better accuracy of pathological information when generating reports.

4.3.2.1.2. Long coherent report. The need to generate longer coherent reports is more commonly seen in ARRG. For this reason, it is necessary to improve the simple CNN-RNN architecture designed for AIC. To achieve this, CNN-MSRNN [55] proposed using three stacked LSTMs to substitute the general RNN decoder. This model performed better in generating reports for normal samples than for abnormal samples. Furthermore, CNN-HRNN-MultiAtt [52] adopted the CNN-HRNN architecture and proposed a multi-step attention mechanism which decomposed the single-step visual attention into the channel- and position-level processing.

Apart from the traditional encoder-decoder architecture, recent studies have explored the use of transformer-based methods. Both MemoryDrivenTR [13] and DSTR [69] implemented a transformer encoder for the secondary encoding of visual features and followed by a transformer decoder for report generation. Moreover, MemoryDrivenTR used a memory mechanism to enhance the transformer decoder, while DSTR utilized extract disease labels to fine-tune the model to improve the clinical coherence of the report. CDGPT2 [62] and ConsecutiveTR [63] directly employed large language models pre-trained using the transformer architecture as the decoders. Additionally, ConsecutiveTR added an intermediate process to perform an abstract transformation from image features to high-level reporting context. The Natural Language Generation (NLG) scores reported in the original studies [13,62,63,69] suggested that ConsecutiveTR and MemoryDrivenTR had similar performance, whereas CDGPT2 and DSTR also had similar performance but worse than the former two.

Other studies have combined RL with the above architectures to address this task. CMAS [61] and CNN-HRNN-RL [59] were based on the CNN-HRNN architecture, while RTMIC [51] used a CNN-transformer architecture. Among them, HRGR-Agent and RTMIC used CIDEr as rewards, while CMAS used BLEU-4 as the basis for rewards. CNN-HRNN-RL also incorporated a novel reward regarding clinical efficacy.

4.3.2.1.3. Utilizing auxiliary classification. To enable the CNN-RNN architecture to be utilized for long report generation, ARRG models often incorporate classifiers alongside the traditional CNN-RNN architecture. For example, TieNet [34] performed disease classification and report generation simultaneously, utilizing two attention mechanisms to highlight essential words and image areas over the outputs. However, it might sacrify the classification performance for better generation performance [65]. Therefore, Vispi [65] proposed to perform disease classification and report generation in order, thereby utilizing the former to enhance the latter. Moreover, Vispi's classification module not only predicted disease labels but also located the lesion areas. Hence, the generation module could separately generate overall abnormal findings, fine-grained abnormal findings, and normal findings. In addition, FCN-MLC-LSTM [71] proposed using U-Net's down-sampling portion [103] as the CNN encoder backbone to identify and classify lesions in mammography. Then the corresponding label was transformed into semantic embedding and passed to the decoder.

Moreover, this approach can also be seen in other architectures. In the CNN-HRNN architecture, CNN-HRNN-CoAtt [12] and CNN-HRNN-AttF [54] performed visual feature extraction and label prediction during the encoding stage, of which the former jointly represented the features and labels through the co-attention mechanism, and the latter directly passed them to separate decoders. In addition to leveraging the detected tags, CNN-HRNN-GLP [60] proposed embedding the outputs of the decoders into the same semantic space and augmenting the training data by using similarity matching. The diversity of the generated sentences was consequently improved. To mitigate data imbalance, CNN-HRNN-Dual [64] proposed dual word-level LSTMs with a sentence predictor, which processed normal findings and abnormal findings, respectively. Notably, CNN-HRNN-GLP introduced a novel pooling approach for its classification module, which achieved higher recall and precision than traditional global feature pooling. The other

involved models are SentSAT-KG [56], which used graph structures to embed prior knowledge into their CNN-HRNN-based model to improve generation performance, where the graph embedding module was pre-trained via multi-label classification; and GAN-ARAE [50], which proposed using a GAN's generator to extract image features and using the decoder of the Adversarially Regularized Autoencoders [104] to generate a diagnosis label and text simultaneously.

4.3.2.1.4. Leveraging report hierarchy. The final method is explicitly designed for chest X-rays to generate the two-section reports (i.e. findings and impression). This method leverages the CNN-HRNN architecture and takes into account the report hierarchy rather than the linguistic hierarchy. In this regard, the findings and impression sections, which indicate the detailed descriptions of images and the corresponding summaries, are usually handled by different modules and trained jointly or separately. In CNN-HRNN-RecAtt [14], the visual features were passed to an RNN decoder to generate the impression. Subsequently, a sequence-to-sequence model was employed to extract the semantic features of impression, which were combined with the regional visual features to generate the findings recurrently. CNN-HRNN-IDC [58] used the same structure as CNN-HRNN-RecAtt. However, the semantic features of impression were combined with visual features as additional constraints to initialize the decoder of the findings module, making the generated sentences conform to the topic of the entire report. Additionally, STS [57] combined the ideas of Vispi and CNN-HRNN-RecAtt. It first utilized a binary classifier to distinguish normal image samples from abnormal cases. A model with CNN-SRNN architecture was then used to generate findings from the classified images. Finally, a summarization module based on sequence-to-sequence architecture was employed to summarize the generated findings into impressions.

4.3.2.2. Structured report generation. This task aims to generate structured reports comprised of a large number of labels that refer to anatomical sites and lesions together with the corresponding intensity, location, shape, size, etc. These detailed descriptive entities can be easily converted into narrative reports. The intuitive solution is to use CNN-based classification framework. For instance, LesaNet [15] and its predecessor, pre-LesaNet [73], used CNNs to predict 171/145 predefined structured report entities. These detailed descriptive entities can be easily converted into narrative reports, such as by ontological mapping used in CNN-FFL [16] or by decision trees used in MultusRadBot [80,87]. CNN-SVM [75] decomposed the task into 43 independent classification questions with close-ended answers and open-ended numerical answers, each associated with its own CNN.

CNN-RNN architecture is also eligible for this task. Cascade CNN-RNN [33] proposed recurrently learning different levels of information through weight reuse. However, such a design was prone to raising error propagation and deteriorating the model generalization ability. To conquer this, Sequence CNN-RNN [53] changed the timing and manner of passing image feature embeddings onto the decoder. Such that the model can maintain the correlation between the image features and the generation process of structured report entities at the time-step level, resulting in a better generation performance. CNN-SAT [68] employed the classic AIC model [102] and proved that introducing additional patient data could effectively increase the percentage of correctly generated structured diagnostic report sentences. In addition to CNN-RNN architecture, CNN-SRNN-Att [88] replaced the 1-layer RNN decoder with two-stacked LSTMs to generate structured report entities, which were then expanded into reports by templates.

On the other side, we found that ARRG models prefer generating structured report entities for lumbar spine MRI reporting [17,79,80,87]. However, the structural correlations of the lumbar spine are an important basis for the reporting, necessitating a segmentation of the disc regions. In this regard, MultusRadBot employed DeepSPINE [105] in their segmentation module, while RGAN-PL [79] and NSL-AGNet [17] applied GAN to segment and classify the lumbar spine structures. The classification results were expanded to structured report entities and compiled to report templates by logical reasoning methods and rule-based methods. Moreover, NSL-AGNet converted the prior knowledge of spinal structures into graph structures and embedded them into the model to improve the consistency of spinal structure identification.

4.3.2.3. Retrieval-based approach and Hybrid retrieval-generation approach. Unlike the traditional generation-based approach, the retrieval-based approach does not generate new text. Instead, it fetches the topmost relevant data from an existing database. Generally, the data are unified into a joint embedding space to compute their similarity, and the results will be stored in a database for later usage. For example, CNN-CVSE [18] proposed a metric learning-based method to learn the visual-semantic embeddings, whereby the fine-grained similarity between the lesion regions and abnormal findings could be measured. However, although it successfully mitigates the weaknesses of the generation-based approach in producing repetitive sentences and bias toward normal findings, its performance appeared not to achieve satisfactory NLG scores. On the other hand, RTEx [19] compared the visual features between the input images and the database images using cosine similarity and assigned the diagnostic sentences of the most similar image to the target image. It achieved high clinical correctness through retrieval constraints on image tags and priority training on abnormal exams.

In addition, several studies were interested in using the retrieval-based approach to complement report generation. One example is HRGR-Agent [20] which uses a CNN-SRNN architecture to combine sentence retrieval and report generation. It utilized RL to train the model with CIDEr as a reward, in which an agent determined whether to use a word decoder to generate sentences or to retrieve directly from the database based on the topic states of a sentence decoder. Another study, KERP [21], utilized graph structures to represent reports as intermediate states in the process of image-to-report generation. These states were then used to perform disease classification and retrieve template sentences. KERP introduced a Graph Transformer to process multi-domain graph structure data and re-write the template sentences into final reports. It outperformed HRGR-Agent in BLEU and ROUGE scores, while its CIDEr score was lower.

4.4. RQ3: evaluation

The ability to generate high-quality radiology reports that are both readable and accurate is the key consideration when developing an effective ARRG model. The included studies employed a wide range of methods to evaluate the outputs of ARRG quantitatively or qualitatively. The quality of automatically generated text can be measured as a function of how closely it resembles a natural language as it is used by humans. A range of automatic measures have been devised to the machine translation field and have been spread to natural language generation (NLG). However, these common NLG metrics are not specifically designed for the ARRG task. Thus, complementary metrics were proposed for measuring the quality of radiology reports with respect to their clinically relevant content. Nonetheless, the finer-grained judgements of the quality of automatically generated text can currently only be provided by humans themselves. Some of the included studies adopted human expert evaluation. Typically, a Likert scale is used to elicit responses from medical experts and various statistics are used to measure the reliability of these responses. The full range of methods used to evaluate ARRG is summarized in Table 4.

4.4.1. Quantitative evaluation

We hoped to compare all ARRG models directly by referring to specific values of the relevant metrics. Unfortunately, the reported results of many ARRG models and other studies' baseline experiments

Table 4
The quantitative and qualitative metrics used for evaluating ARRG.

Metrics	Description	Used by	Count
Quantitative evaluation			
BLEU [106]	A precision-based metric that counts how many n-grams from the generated text exist among ground truth references.	[11–14,18–21,33,34,50–65,68,69,71,81,83,88,107].	33
ROUGE [108]	A recall-based metric that counts how many n-grams from the ground truth references exist in the generated text.	[11–14,18,20,21,34,50,52,54–65,69,81,83,107]	26
METEOR [109]	A metric that counts unigram matches using an F1-like measure.	[11–14,18,34,50,54,55,57–60,62–64,69,81,107]	19
CIDEr [109]	The cosine similarity between the vectors weighted by term frequency-inverse document frequency, which measures the consensus between the ground truth references and the generated text.	[11,12,20,21,50–52,55–62,64,65,69,71,81]	20
Keywords Accuracy [14]	The ratio of the number of diagnostic keywords in the generated reports to the number of all diagnostic keywords among the ground truth references.	[14]	1
Clinical Efficacy [59]	Measures the accuracy, precision, and recall of disease labels extracted by CheXpert from the ground truth references and the generated reports.	[13,59,63,69]	4
MeSH Accuracy [52]	The ratio of the number of MeSH terms correctly generated by a model to the number of all MeSH terms in the ground truth references.	[52]	1
Anatomical Relevance Score [83]	Matches the words in GRs against the terminology of the anatomical class of interest.	[83]	1
Medical Abnormality Terminology Detection Accuracy [20]	Compares the average precision and average false positives of 10 most frequent medical abnormality terminologies in the ground truth references and the generated reports.	[20]	1
MIQRI [56]	Evaluates the quality of paired reports by graph matching.	[56]	1
Qualitative evaluation			
Human expert evaluation	E.g. average score, average preference percentage, etc.	[20,21,62,88]	4
Likert scale [110]	A rating system used to measure the opinion of medical experts regarding the quality of generated reports.	[83]	1
Cohen's kappa coefficient [47]	A statistic used to measure inter-rater reliability.	[19,87]	2
Fleiss' kappa coefficient [111]	A statistic used to measure inter-rater reliability.	[50]	1
Cronbach's alpha coefficient [112]	A measure of internal consistency.	[87]	1

proved inconsistent. First, there is no benchmark that would allow the models to be evaluated on a common dataset. Second, different metrics are used to report the results. To provide an overview of the performance, we restricted the comparison between models to the baseline experiments of a single study and exhibited them in a performance matrix, as shown in Fig. 10. More specifically, we designed a simple metric score merging algorithm that uses percentages to show the performance gap between models. Given any target model t and baseline model b, the performance gap $p(t,b)$:

$$\mathrm{p} = \frac{y_t - y_b}{y_b}, \# \tag{1}$$

$$y = \frac{\sum m_i}{num(i)}, \# \tag{2}$$

where $i \in BLEU, ROUGE, METEOR, CIDEr$, m_i is a specific metric value, and $num(i)$ is the number of the metrics used. If $i = BLEU$, then:

$$m_i = \frac{\sum m_{BLEU-x}}{num(x)}, (x \in 1,2,3,4).\# \tag{3}$$

Note that for each row in Fig. 10, the red colour with a positive value indicates that the target model outperformed the baseline model and vice versa. From the perspective of columns, the target models are compared by benchmarking to the same baseline model. However, such ranking results vary with the choice of different baseline models, which also reflects the inconsistency issue.

4.4.2. Qualitative evaluation

We reviewed eight relevant studies that have presented human expert evaluation. Although some studies claimed that their models can generate more accurate and reasonable reports than baseline models, the gap between these generated reports and expert-written reports was not clearly evaluated [19–21]. In contrast, several studies highlighted the need for further advancement of ARRG models for more reliable outputs [62,80,83,87]. In particular, CDGPT2 [62] pointed out that their model was capable of generating correct reports for 99% of normal samples. However, the generated reports for abnormal samples suffer from missing information and incorrect diagnosis and often lack the necessary details to describe the abnormalities present in the images. In CNN-SRNN-Att [88], the evaluation noted that clinicians preferred the combination of visual interpretation and "human style" textual explanations for pelvic X-rays. The reliability assessment of MultusRadBot [80,87] disclosed a small "opinion discrepancy" between the generated reports and expert-written reports in lumber spine MRI, and such discrepancies in reporting can significantly impact subsequent clinical decisions. In GAN-ARAE [50], the model's usefulness was confirmed in helping radiologists achieve higher accuracy diagnoses and, perhaps, quicker decision-making processes for edge samples.

5. Discussions

5.1. Current state of affairs and challenges

Radiology images, especially ultrasound images, have relatively low resolution and blurred boundaries between the foreground and background. Radiology reports, on the other hand, tend to be lengthy, complex, and heterogeneous, covering descriptions of findings, impressions, and other patient-related information. They also contain expressions that convey negation and uncertainty. Furthermore, the structure and style of radiology reports may vary significantly between institutions or individual radiologists, raising the concerns of interobserver variability in the training data. During the clinical reporting process, the radiologist's wording could have been influenced by affective (unconscious emotional reaction) and cognitive (distortions of thinking) biases [87]. When the data originate from a small number of institutions, they may not be representative, which may lead to overfitting [116]. Moreover, the available open-source datasets are often limited in size and unbalanced in the distribution of normal and abnormal samples, making it even more difficult to train a robust model. Across all datasets identified in this review, we argue that only the

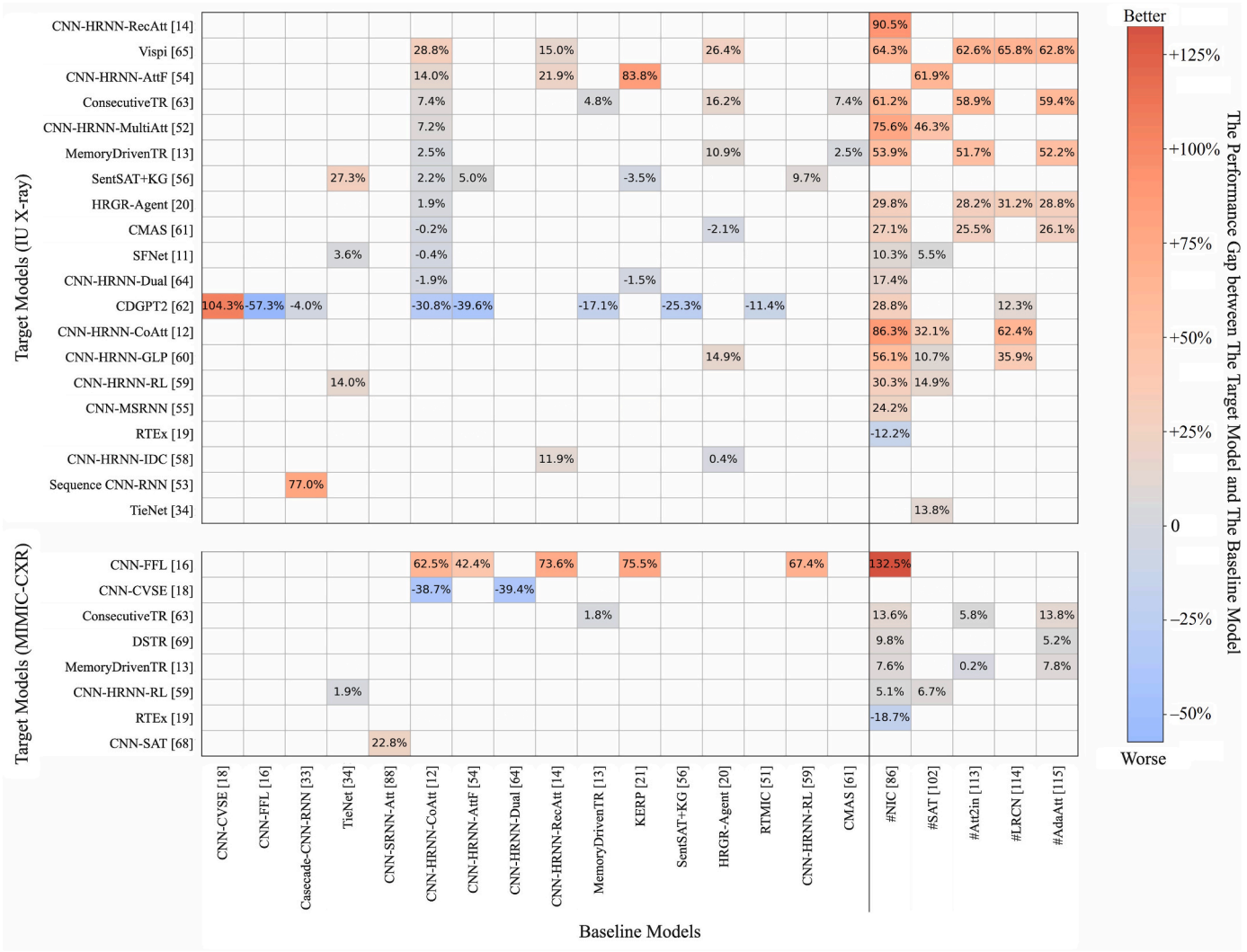


Fig. 10. Model performance matrix. Each row is a set of comparisons of model performance gaps computed based on the baseline experiments of the target model. Results are limited to the IU X-ray dataset (top) and the MIMIC-CXR dataset (bottom). Some common AIC baseline models are used in addition to the ARRG models, including #NIC [86], #SAT [102], #Att2in Ref. [113], #LRCN [114] and #AdaAtt [115]. The target models are sorted in descending according to the performance compared to the CNN-HRNN-CoAtt [12] and #NIC models.

MIMIC-CXR dataset meets three key conditions: large-scale, publicly available and containing original reports.

Regarding the assessment of the generated reports, many ARRG models were benchmarked using ordinary AIC metrics, such as BLEU, ROUGE, METEOR and CIDEr, which are based on n-gram overlap and focus more on language fluency. However, n-gram overlap is neither necessary nor sufficient for two sentences to convey the same meaning [117]. It was widely believed that accurate detection of pathology should take precedence over language fluency when evaluating the generated reports. Hence various clinical efficacy metrics based on the accuracy, precision, and recall of disease labels were designed for the ARRG task [14,20,52,59,83]. Nevertheless, these metrics might not be sufficient for evaluating a report since the pathological description not only concerns specific disease labels but also involve their qualifiers, certainties and negations. Although MIQRI [7] have taken into consideration all these attributes, its effectiveness is yet to be fully proven.

Due to the discrepancy between general image/caption datasets and radiology image/report datasets, using conventional AIC models on the ARRG task might only produce reports that look real but are not clinically correct. Therefore, it is necessary to tailor DL approaches specifically for the ARRG task. To summarize, retrieval-based methods leverage the similarity of data features, which might generate fewer repetitive sentences and mitigate the bias toward generating normal findings [18–21]. RL can train an ARRG model to generate reports toward specific metrics of interest, such as better pathological accuracy [59]. Some studies enhanced the generating process by integrating auxiliary classifiers [11,12,21,34,54,56,57,60,64,65] or taking into account the report hierarchy, while other studies substituted the conventional generation process by generating more informative structured report entities [15–17,33,53,68,73,75,79,80,87,88]. In order to provide radiologists with more interpretable information, researchers offered many ideas, including applying GAN to generate similar images [50], employing class activation maps and saliency maps (e.g. attention maps) to bring visual interpretation [12,13,18,21,34,50,59,62,65,68,88], or using the predicted pathology labels from the auxiliary classifier as supplementary of the generated reports. It has also been demonstrated that introducing patient background information could have positive impacts [52,68]. There were also studies that proposed separate processing of normal and abnormal samples to address the data imbalance problem [57,64,65].

It can also be observed that most of the ARRG systems were designed for X-ray tests, in which the difficulty of data access might be the leading cause of why researchers preferred to show the results on the IU X-ray than the larger MIMIC-CXR dataset. For CT scans, the proposed ARRG

systems were largely based on image classification. Current MRI datasets were all focused on the lumbar spine. Due to the structural correlation of the lumbar spine and the particular report format, the proposed systems tend to deem it as segmentation and classification problems and combine with non-DL approaches to compile the predicted labels into reports. Considering that both US and X-ray datasets are intrinsically similar, even though US reports tend to be shorter, the corresponding ARRG systems could be easily repurposed.

5.2. Future research Directions

ARRG systems certainly have the potential to streamline clinical workflow. However, the current state of the art in ARRG has yet to generate high-quality reports. We propose exploring the following aspects for further improvement.

First, language models pre-trained on large datasets can help alleviate the data scarcity issue by allowing ARRG models to use smaller training datasets for fine-tuning. In addition, the capability of large language models to capture semantically equivalent contexts can help develop better evaluation metrics. Despite their great success in a wide range of NLP tasks [118,119], their full potential in the ARRG domain is yet to be realized due to the delay of technique shift across domains. Therefore, together with large language models, which are commonly trained by transformer architectures, we can expect ARRG to follow a trend observed in NLP by shifting away from RNNs to transformers.

Second, the automatically generated text sometimes deviates from the ground-truth text in terms of structure, coverage, and lexical content. This can be attributed to the interobserver variability within the training data. This variability is caused by the inherent diversity of natural language, which allows for the same information to be expressed in numerous ways. These issues can be mitigated by adopting structured reports, which use uniform language and structure to describe radiology findings accurately [120]. Even though structured reporting is increasingly being used, especially in abdominal and neuroradiological CT and MRI reports [121], the cultural and technological shift required will inevitably delay their widespread adoption. Instead, we propose resorting to NLP approaches to automatically structure the legacy radiology reports and used them not only to reduce interobserver variability, but also to train DL approaches to automatically generate structured reports.

Third, the metrics for evaluating the generated radiology reports require further investigation as the current metrics cannot comprehensively assess the quality of the generated report. Advances in measuring the semantic similarity for general image description can be seen in SPICE [122], which is a new concept-based AIC metric using a semantic graph to capture the meaning of two captions. This measurement approach has been widely adopted in recent AIC models. Although MIQRI made the first attempt in ARRG to develop a graph-based clinical semantic measurement, its effectiveness is underdetermined. We believe future research can draw inspiration from SPICE and MIQRI to develop an evaluation system that can capture the correctness of pathological information and the relationship between pathological attributes and thoroughly verify its effectiveness via thorough baseline experiments and manual evaluation. Finally, radiology images other than chest X-rays are rarely available publicly and at scale. Therefore, appropriate domain-specific evaluation metrics together with large-scale publicly available datasets are required to both deepen and broaden the existing research into ARRG.

6. Conclusions

In this study, we review of 41 ARRG studies selected from 1443 records. We first illustrate six publicly available radiology image/report datasets and other auxiliary datasets involved in these studies. Secondly, we provide an overview of the frameworks, network architectures, and techniques employed by ARRG models. Each individual model is discussed from the perspective of its targeted tasks and implementation methods. Then, we summarize the evaluation methods used in these models. Meanwhile, we illustrate the relative performance of these models in quantitative evaluation and present the opinions of human experts to the automatically generated reports. Finally, we discuss the current challenges and future research directions of ARRG. Overall, ARRG is promising in playing an important role in the clinical workflow, but further advancement is still required to generate reports that are indistinguishable from those written by radiologists.

Author statement

No human subjects were involved in this study.

Summary

- Automatic radiology report generation is more challenging than conventional image captioning due to some inherent properties of the data.
- Natural language processing can be used to automatically structure the legacy radiology reports and to reduce the sources of variation and bias in the training data.
- Language models that are pre-trained on large datasets may enable deep learning approaches to be fine-tuned automatic radiology report generation using relatively small training datasets.

Declaration of competing interest

We declare that we have no conflict of interest.

Acknowledgements

This work is part of a PhD project funded by China Scholarship Council-Cardiff University Scholarship. The scholarship has been awarded to Y.L. The project is supervised by I.S. and H.L.

APPENDIX

A. Comparison of ARRG Systems

Method (Proposed by)	Field	Key Technique	Problem Representation	Pre-trained model
CNN-based ARRG systems				
CNN-CVSE (Ni et al., 2020) [18]	X-ray: chest	– Image feature extraction: DenseNet-121 [123] – Semantic embedding extraction: BioWordVec [124]	– Sentence selection (find the closest abnormal findings from database via visual-semantic similarity) – Heatmap on image (attention map)	– Pre-trained DenseNet-121 [123] – Pre-trained BioWordVec [124]

(continued on next page)

(*continued*)

Method (Proposed by)	Field	Key Technique	Problem Representation	Pre-trained model
		– Joint embedding and visual-semantic similarity: conditional visual-semantic embeddings		
RTEx (Kougia et al., 2021) [19]	X-ray: chest	– Ranking: RTEx@R: CNN (DenseNet-121 [123]) – Tagging: RTEx@T: CNN (DenseNet-121 [123]) – Captioning: RTEx@X: CNN (DenseNet-121 [123])	Sentence selection (find the most similar image in database and assign its text to the target image)	None declared
CNN-FFL (Syeda-Mahmood et al., 2020) [16]	X-ray: chest	– Define Fine Finding Labels (FFL) – Extract FFL from reports and construct dataset – FFL prediction: CNN (FPN [107] that combines pre-trained VGG-16 [125] and ResNet-50 [126]) with dilated blocks – Ontological mapping	Structured report generation (fine-grained finding labels)	– Pre-trained VGG-16 [125] and ResNet-50 [126] on ImageNet [127]
MultusRadBot (Lewandrowski et al., 2020) [80,87]	MRI: lumbar spine	– Image processing: CNNs (FC-DenseNet [128] and DeepSPINE [105]) and principal component analysis (PCA) – Multi-class classification: two CNNs (VGG [125]) – Report generation: decision trees	– First: segment, stacked region cropping, and diameter prediction – Then: multi-label classification – Finally: compile to reports	None declared
Pre-LesaNet (Yan et al., 2019) [73]	CT: 115 body parts, 27 types, and 29 attributes	– Lesion annotation: CNN (VGG-16 [125])	Structured report generation (multi-label classification of 145 labels)	None declared
LesaNet (Yan et al., 2019) [15]	CT: 115 body parts, 27 types, and 29 attributes	– Image feature: CNN (VGG-16 [125]) – Classifier: a score propagation layer – Enrich label relations: · Label expansion ·Relational hard example mining	Structured report generation (multi-label classification of 171 labels)	None declared
CNN-SVM (Loveymi et al., 2021) [75]	CT: liver	For each question: – Image feature: CNN (MobileNet [129], LeNet-5 [130]) – Classifier: SVM with linear kernel	Structured report generation (independent classification for 43 pre-defined questions with close-ended answers and open-ended numerical answers).	MobileNet [129] that trained on ImageNet [127]
CNN-RNN-based ARRG systems				
Cascade CNN-RNN (Shin et al., 2016) [33]	X-ray: chest	– Image representation: CNN (GoogLeNet [131]) – Text generation: LSTM [89]/GRU [90] – Training method: re-train the model in a cascade workflow	Structured report generation (composite image labelling)	None declared
Sequence CNN-RNN (Gasimova, 2020) [53]	X-ray: chest	– Enriched concept extraction (multi-label classification: CNN – Image representation: CNN (VGG-16 [125]/ResNet-50 [126]) – Text generation: LSTM [89]	Structured report generation (visually significant medical concepts extracted from raw reports)	Pre-trained VGG-16 [125]/ResNet-50 [126] that trained on ImageNet [127]
CNN-SAT (Rodin et al., 2019) [68]	X-ray: chest	– Image encoder: CNN (DenseNet-121 [123]) – Text decoder: LSTM [89] with a soft attention [101] – Additional background information embedding	– Structured report generation (assertion pair about pathology presence, location and severity) – Heatmap on images (attention map)	– DenseNet-121 [123] that trained on the ChestX-ray14 [77] – Word2vec that trained on PubMed [132]
TieNet (Wang et al., 2018) [34]	X-ray: chest	– Image encoder: CNN (ResNet-50 [126]) – Text decoder: LSTM [89] with soft attention [101] and self-attention mechanism proposed by Lin et al. [ref]) – Joint learning of visual and textual weighted features for classification	– Multi-label disease classification – Report generation (findings + impression) – Heatmap on text (attention map)	– ResNet-50 [126] that trained on ImageNet [127] – The Gensim word2vec implementation [133] trained on PubMed articles
Vispi (Li et al., 2019) [65]	X-ray: chest	– Classification: CNN (DenseNet-121 [123]) – Localization: CNN (Grad-CAMs [134]) – Generation: attention-based encoder-decoder model ·Encoder: CNN (ResNet-101 [126]) ·Decoder: LSTMs [89] with soft attention [101]	– Multi-label disease classification and localization – Report generation (findings + impression) – Heatmap on image (class activation map)	– ResNet-101 [126] that trained on ImageNet [127] – DenseNet-121 [123] that trained on ImageNet [127] and ChestX-ray8 [77]
FCN-MLC-LSTM (Sun et al., 2019) [71]	Mammography	– Image encoder and multi-label classification: · U-net's down-sampling part [103]	– Multi-label disease classification – Report generation	None declared

(*continued on next page*)

(*continued*)

Method (Proposed by)	Field	Key Technique	Problem Representation	Pre-trained model
		· Fully Convolutional Network based on ResNet [126] · Multi-label classification (MLC): the last two layers of ResNet [126] – Text generation: LSTM [89] – Post-processing: beam search		
CNN-RNN-Merging (Alsharid et al., 2019) [83]	US: fetal	– Image representation: CNN (VGG-16 [125]). – Text representation: LSTM [89]/GRU [90] – Joint representation: merging the output layer of CNN–FCN and the hidden layer of RNN.	Report generation (voice-over caption from sonographer)	– VGG-16 [125] that trained on ImageNet [127] – The Google Code word2vec implementation that trained on Google News dataset [135], GloVe that trained on Wikipedia-2014 corpus [136].
HCNN-RNN (Zeng et al., 2018) [82]	US: gallbladder, liver, kidney	– Organ (coarse) classification: CNN (VGG-16 [125]) – Disease (fine-grained) encoder: CNNs (VGG-16 [125]) – Text decoder: LSTM [89]	Report generation (short annotation text that explain the disease information in the ultrasound image)	VGG-16 [125] that trained on ImageNet [127]
FRCNN-RNN (Zeng et al., 2020) [81]	US: gallbladder, kidney, liver	– Image encoder: VGG-16 [125] and Faster RCNN [137] – Text decoder: LSTM [89]	– Image detection – Report generation (a short description of ultrasound image)	Pre-trained VGG-16 [125]
SFNet (Zeng et al., 2020) [11]	– US: gallbladder, kidney, liver – X-ray: Chest	– Image detection: CNNs (ResNet-50 [126], Faster RCNN [137]) – Text: LSTM with a sentinel gate [138]	– Lesion area detection and pathological information classification – Report generation (impression)	ResNet-50 [126] that trained on ImageNet [127]
CNN-SRNN-based ARRG systems				
CNN-MSRNN (Singh et al., 2019) [55]	X-ray: chest	– Image encoder: CNN (Inception-v3 [139]) – Text decoder: multi-stage stacked LSTMs initialized by RadGlove [140]	Report generation (findings + impression)	– Inception-v3 [139] that trained on ImageNet [127] – Pre-trained GloVe [136], RadGlove that trained on 4.5 million radiology reports [140]
CNN-SRNN-Att (Gale et al., 2019) [88]	X-ray: hip fractures in pelvic	– Image representation: CNN (DenseNet [123]) – Text generation: two stacked LSTMs with soft attention [101]	– Structured report generation (each sentence had a general structure) – Heatmap on images (saliency map)	Pre-trained DenseNet [123]
CNN-HRNN-based ARRG systems				
CNN-HRNN-CoAtt (Jing et al., 2018) [12]	– X-ray: chest – Mixed: 21 different sub-categories (from PEIR)	– Image encoder and tag prediction: CNN (VGG-19 [125]) – Sentence encoder: LSTM [89] with proposed co-attention mechanism – Word decoder: LSTM [89]	– Multi-label classification (tags prediction) – Report generation (findings + impression) – Heatmap on image and text (co-attention map)	None declared
CNN-HRNN-AttF (Yuan et al., 2019) [54]	X-ray: chest	– Image encoder: a multi-view CNN (ResNet-152 [126]) – Medical concepts prediction – Sentence decoder: LSTM [89] with soft attention [101] – Word decoder: concept enriched LSTM with attention mechanism (soft attention [101])	– Multi-label classification – Report generation (findings + impression) – Heatmap on image (attention map)	ResNet-152 [126] that pre-trained on CheXpert [78]
CNN-HRNN-MultiAtt (Huang et al., 2019) [52]	X-ray: chest	– Image encoder: CNN (ResNet-152 [126]) – Sentence decoder: LSTM [89] with proposed multi-attention mechanism – Word decoder: Bi-LSTM [141] with embedded background information and soft attention [101]	Report generation (Findings + impression)	– ResNet-152 [126] that trained on ImageNet [127] – Pre-trained GloVe [136]
CNN-HRNN-GLP (Yin et al., 2019) [60]	X-ray: chest	– Multi-label classification: CNN (DenseNet [123]) with a global label pooling mechanism – Sentence encoder: LSTM [89] with topic attention mechanism (which closely resembles co-attention [12]) – Word decoder: LSTM [89]	– Multi-label classification (abnormality detection) – Report generation (findings + impression)	DenseNet [123] that trained on ImageNet [127], and then pre-trained it on IU X-rays [49]
CNN-HRNN-Dual (Harzig et al., 2019) [64]	X-ray: chest	– Image encoder: CNN (ResNet-152 [126]) – MTI (Medial Text Indexer) tag prediction – Sentence encoder: LSTM [89] with soft attention [101] – Word decoder: an abnormal word LSTM and a normal word LSTM	– Multi-label classification (MTI tags) – Report generation (findings + impression)	Word2vec that trained on PubMed and Wikipedia [132]
CNN-HRNN-RecAtt (Xue et al., 2018) [14]	X-ray: chest	– Image encoder: CNN (ResNet-152 [126]) – Sentence generative model: LSTM [89]	– Report generation (impression) – Report generation (findings)	ResNet-152 [126] that trained on ImageNet [127]

(*continued on next page*)

(*continued*)

Method (Proposed by)	Field	Key Technique	Problem Representation	Pre-trained model
		– Recurrent paragraph generative model: · Sentence encoder: Bi-LSTM [141]/1D-CNN [142] · Sentence decoder: stacked 2-layer LSTM with soft attention [101]		
CNN-HRNN-IDC (Xue et al., 2019) [58]	X-ray: chest	– Image encoder: CNN (ResNet-152 [126]) – Impression generation: LSTM [89] – Findings generation: · Sentence encoder: Bi-LSTM [141] · Sentence decoder: two stacked LSTMs with soft attention [101]	– Report generation (impression) – Report generation (findings)	ResNet-152 [126] that trained on ImageNet [127]
STS (Singh et al., 2021) [57]	X-ray: chest	– Image classification module: CNN (Inception-v3 [139]) – Generation module: CNN-SRNN (Inception-v3 [139] as encoder and stacked LSTM as decoder [89])) – Summarization module: proposed by [143] · Encoder: Bi-LSTM [141] with soft attention [101] · Decoder: LSTM [89] with "copy" mechanism [ref]	– Image classification – Report generation (findings) – Report summarization (impression)	– Inception-v3 [139] that trained on ImageNet [127] – Pre-trained Glove [136], pre-trained RadGlove [140]
Transformer-based ARRG systems				
MemoryDrivenTR (Chen et al., 2020) [13]	X-ray: chest	– Visual extractor: CNN (ResNet-101 [126]) – Encoder: the transformer's encoder [94] – Decoder: the transformer's decoder [94] with memory module (a relational memory and a novel memory-driven conditional layer normalization)	– Report generation (long report) – Heatmap on image (attention map)	ResNet-101 [126] that trained on ImageNet [127]
DSTR (Lovelace et al., 2020) [69]	X-ray: chest	– Image encoder: CNN (DenseNet-121 [123]) – Report generation: transformer [94] with differentiable sampling – A differentiable CheXpert model [78] to fine-tune the generation model: · Bi-LSTM [141] with soft attention [101]; · or CNN with scaled dot-product attention [94]	– Report generation (findings)	– Pretrained DenseNet-121 [123] – Word2Vec [144] that pre-trained on MIMIC-CXR [66]
CDGPT2 (Alfarghaly et al., 2021) [62]	X-ray: chest	– Image encoder: · visual feature: CNN (CheXNet [145]) · semantic features: word embeddings (McDonald's word2vec implementation [146]) – Text Decoder: transformer (DistilGPT2 [147]) – A conditioning approach	– Report generation (findings + impression) – Heatmap on image (attention map)	– CheXNet [145] that trained on ChestX-ray14 [77] – Word2vec that trained on MEDLINE/PubMed data [146] – DistilGPT2 [147] that trained on OpenWebTextCorpus [148]
ConsecutiveTR (Nooralahzadeh et al., 2021) [63]	X-ray: chest	– A visual backbone (CNN: DenseNet [123]) – A visual language model for high-level context generation: Meshed-Memory Transformer [149]) – A language model for narrative report generation: transformer-based encoder-decoder model (BART [150])	– High-level context (resemble structured report entities) – Report generation (full report)	– Pre-trained DenseNet [123] – Pre-trained language model BART [150]
RL-based ARRG systems				
HRGR-Agent (Li et al., 2018) [20]	X-ray: chest	– Image encoder: CNN (DenseNet [123]/VGG-19 [125]) – Sentence decoder: stacked RNNs with a soft attention varient [ref] – A generation module: same as the sentence decoder – Training method: hierarchical reinforcement learning using a CIDEr based reward	Hybrid report generation and sentence selection (findings)	DenseNet [123] that jointly pre-trained on ChestX-ray8 [77] and CX-CHR [20]
CNN-HRNN-RL (Liu et al., 2019) [59]	X-ray: chest	– Image encoder: CNN (DenseNet-121 [123]) – Sentence decoder: LSTM [89] – Word decoder: LSTM [89] with attention mechanism	– Report generation (findings) – Heatmap on images (attention map)	Word embedding that pretrained with Gensim [151]

(*continued on next page*)

(*continued*)

Method (Proposed by)	Field	Key Technique	Problem Representation	Pre-trained model
		– Training method: reinforcement learning using both language fluency and clinically coherent reward		
CMAS (Jing et al., 2019) [61]	X-ray: chest	– Image encoder: CNN (ResNet-50 [126]) – Global state encoder: LSTM [89] with soft attention [101] – Planner: a two-layer feed-forward network – Writer: · Normality Writer: LSTM [89] · Abnormality Writer: LSTM [89] – Training method: reinforcement learning using a reward based on BLEU-4 – Duplicated modules for *Findings* and *Impression.*	– Report generation (findings + impression)	ResNet-50 [126] that trained on ImageNet [127]
RTMIC (Xiong et al., 2019) [51]	X-ray: chest	– Image encoder: a bottom-up region detector (CheXNet [145]), a top-down visual encoder (the transformer's encoder [94]) – Captioning decoder: the transformer's decoder [94] – Training method: self-critical reinforcement learning using CIDEr as reward	– Image captioning (findings)	CheXNet [145] that trained on ChestX-ray14 [77]
GAN-based ARRG systems				
GAN-ARAE (Spinks et al., 2019) [50]	X-ray: chest	– Text representation: Adversarially Regularized Autoencoders (ARAE) [104], with LSTMs [89] as encoder and decoder. – Image representation: CNN, inverse mapping of GAN's generator – Text generation: ARAE's decoder – Training method: text-to-image GAN (StackGAN [152]) – Visualization: saliency maps from the activation-based attention schemes [153]	– Classification + Image captioning (generate diagnosis label and findings simultaneously) – Similar image generation – Heatmap on images (saliency map)	None declared
RGAN-PL (Han et al., 2018) [79]	MRI: lumbar spine	– Segmentation, and classification: recurrent GAN · Generative network: A deep atrous convolution autoencoder module, with a spatial LSTM [89] in the middle · Discriminative network: an adversarial module – Positional labeling: strong prior knowledge based unsupervised symbolic program synthesis approach – Template-based captioning: symbolic template based structural captioning method	– Segmentation – Classification – Structured report generation	None declared
Graph-based ARRG systems				
KERP (Li et al., 2019) [21]	X-ray: chest	– Encode: visual feature to knowledge graph · CNN + Graph Transformer – Retrieve: knowledge graph to template sequence · Graph Transformer – Paraphrase: template sequence to report · Graph Transformer – Multi-label classification: graph to graph · Graph Transformer	– Multi-label classification – Report generation (findings) – A generated disease graph as concluding information for medical reports – Heatmap on images (attention map)	Extract the visual features from a DenseNet [123] that jointly pre-trained on CX-CHR [20] and ChestX-ray8 [77]
SentSAT-KG (Zhang et al., 2020) [56]	X-ray: chest	– Image encoder: CNN (DenseNet-121 [123]) + GNN (GCN [154]) with soft attention [101] – Sentence decoder: LSTM [89] with soft attention [101] – Word decoder: LSTM [89]	– Multi-label classification – Report generation (findings + impression)	DenseNet-121 [123] that pre-trained on CheXpert [78]
NSL-AGNet (Han et al., 2021) [17]	MRI: lumbar spine	– Image segmentation and classification: GAN with a symbolic graph reasoning module and local semantic attention	– Segmentation and classification of spinal structures – Induction of the pathological relations	None declared

(*continued on next page*)

(*continued*)

Method (Proposed by)	Field	Key Technique	Problem Representation	Pre-trained model
		– Discovery: symbolic logic reasoning for structured report entities (first-order logic programming, meta-interpretive learning) – Report Generation: ruled-based approach	– Structured report generation	

References

[1] Nhs England, Nhs Improvement. Diagnostic imaging dataset statistical release. 2020. https://www.england.nhs.uk/statistics/wp-content/uploads/sites/2/2020/01/Provisional-Monthly-Diagnostic-Imaging-Dataset-Statistics-2020-01-23.pdf. [Accessed 21 February 2022].

[2] Royal College of Radiologists. Clinical radiology UK workforce census 2020 report. 2020. https://www.rcr.ac.uk/system/files/publication/field_publication_files/clinical-radiology-uk-workforce-census-2020-report.pdf. [Accessed 21 February 2022].

[3] Mezrich J. Radiology bottlenecks - is there any utility of just in time inventory principles in the ed setting? Clin Imag 2020;66:54–6. https://doi.org/10.1016/j.clinimag.2020.04.035.

[4] Babu AS, Brooks ML. The malpractice liability of radiology reports: minimizing the risk. Radiographics 2015;35(2):547–54. https://doi.org/10.1148/rg.352140046.

[5] Berlin L. Defending the "missed" radiographic diagnosis. Am J Roentgenol 2001;176(2):317–22.

[6] Berlin L, Hendrix RW. Perceptual errors and negligence. AJR Am J Roentgenol 1998;170(4):863–7.

[7] Kaur N, Mittal A, Singh G. Methods for automatic generation of radiological reports of chest radiographs: a comprehensive survey. Multimed Tool Appl 2021. https://doi.org/10.1007/s11042-021-11272-6.

[8] European Society of Radiology (ESR). Good practice for radiological reporting. Guidelines from the European society of radiology (esr). Insights Imag 2011;2(2):93–6. https://doi.org/10.1007/s13244-011-0066-7 [published Online First: 2011/02/06].

[9] Rimmer A. Radiologist shortage leaves patient care at riskvol. 359. Warns Royal College. Bmj; 2017. p. j4683. https://doi.org/10.1136/bmj.j4683 [published Online First: 20171011].

[10] Ming Y, Hu N, Fan C, Feng F, Zhou J, Yu H. Visuals to text: a comprehensive review on automatic image captioning. IEEE/CAA J Automat Sinica 2022;9(8):1339. https://doi.org/10.1109/jas.2022.105734.

[11] Zeng X, Wen L, Xu Y, Ji C. Generating diagnostic report for medical image by high-middle-level visual information incorporation on double deep learning models. Comput Methods Progr Biomed 2020;197:105700. https://doi.org/10.1016/j.cmpb.2020.105700 [published Online First: 20200811].

[12] Jing B, Xie P, Xing EP. On the automatic generation of medical imaging reports. In: 56th annual meeting of the association for computational linguistics, ACL 2018. Association for Computational Linguistics (ACL); 2018.

[13] Chen Z, Song Y, Chang T-H, Wan X. Generating radiology reports via memory-driven transformer. In: Proceedings of the 2020 conference on empirical methods in Natural Language Processing. EMNLP; 2020.

[14] Xue Y, Xu T, Long LR, Xue ZY, Antani S, Thoma GR, Huang XL. Multimodal recurrent model with attention for automated radiology report generation. Med Imag Comput Comput Assisted Intervent - Miccai 2018;11070(Pt I):457–66. https://doi.org/10.1007/978-3-030-00928-1_52. 2018.

[15] Yan K, Peng YF, Sandfort V, Bagheri M, Lu ZY, Summers RM. Holistic and comprehensive annotation of clinically significant findings on diverse ct images: learning from radiology reports and label ontology. In: 32nd IEEE/CVF conference on computer vision and pattern recognition (CVPR). Long Beach, CA: Ieee; 2019 Jun 16-20.

[16] Syeda-Mahmood T, Wong KCL, Gur Y, Wu JT, Jadhav A, Kashyap S, Karargyris A, Pillai A, Sharma A, Syed AB, Boyko O, Moradi M. Chest X-ray report generation through fine-grained label learning. Medical image computing and computer assisted intervention – miccai 2020. Cham: Springer International Publishing; 2020. 2020//.

[17] 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. https://doi.org/10.1016/j.media.2020.101872 [published Online First: 20201021].

[18] Ni J, Hsu C-N, Gentili A, McAuley J. Learning visual-semantic embeddings for reporting abnormal findings on chest X-rays. Findings of the association for computational linguistics: emnlp 2020. Association for Computational Linguistics; 2020 nov.

[19] Kougia V, Pavlopoulos J, Papapetrou P, Gordon M. Rtex: a novel framework for ranking, tagging, and explanatory diagnostic captioning of radiography exams. J Am Med Inf Assoc 2021;28(8):1651–9. https://doi.org/10.1093/jamia/ocab046.

[20] Li CY, Liang XD, Hu ZT, Xing EP. Hybrid retrieval-generation reinforced agent for medical image report generation. Adv Neurol 2018:31.

[21] Li CY, Liang XD, Hu ZT, Xing EP. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In: Thirty-third aaai conference on artificial intelligence/thirty-first innovative applications of artificial intelligence conference/ninth aaai symposium on educational advances in artificial intelligence; 2019. p. 6666–73.

[22] Ganeshan D, Duong P-AT, Probyn L, Lenchik L, McArthur TA, Retrouvey M, Ghobadi EH, Desouches SL, Pastel D, Francis IR. Structured reporting in radiology. Acad Radiol 2018;25(1):66–73. https://doi.org/10.1016/j.acra.2017.08.005.

[23] Nobel JM, Kok EM, Robben SGF. Redefining the structure of structured reporting in radiology. Insights Imag 2020;11(1):10. https://doi.org/10.1186/s13244-019-0831-6.

[24] Radiological Society of North America. Radreport.Org. https://radreport.org/. [Accessed 21 February 2023].

[25] Spasic I, Zhao B, Jones CB, Button K. Kneetex: an ontology–driven system for information extraction from mri reports. J Biomed Semant 2015;6(1):34. https://doi.org/10.1186/s13326-015-0033-1.

[26] Hassanpour S, Langlotz CP. Information extraction from multi-institutional radiology reports. Artif Intell Med 2016;66:29–39. https://doi.org/10.1016/j.artmed.2015.09.007.

[27] Rubin DL, Kahn J Charles E. Common data elements in radiology. Radiology 2017;283(3):837–44. https://doi.org/10.1148/radiol.2016161553.

[28] Radiological Society of North America. Radelement.Org. https://www.radelement.org/. [Accessed 21 February 2023].

[29] Allaouzi I, Ahmed M Ben, Benamrou B, Ouardouz M. Automatic caption generation for medical images. In: Proceedings of the 3rd international conference on smart city applications (Sca'18); 2018. https://doi.org/10.1145/3286606.3286863.

[30] Monshi MMA, Poon J, Chung V. Deep learning in generating radiology reports: a survey. Artif Intell Med 2020;106:101878. https://doi.org/10.1016/j.artmed.2020.101878 [published Online First: 20200515].

[31] Pavlopoulos J, Kougia V, Androutsopoulos I. A survey on biomedical image captioning. Minneapolis, Minnesota: Association for Computational Linguistics; 2019 jun.

[32] Shin H-C, 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. 2015. https://ui.adsabs.harvard.edu/abs/2015arXiv150500670S. [Accessed 1 May 2015].

[33] 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: 2016 IEEE conference on computer vision and pattern recognition (CVPR). Seattle, WA: Ieee; 2016 Jun 27-30.

[34] Wang XS, Peng YF, Lu L, Lu ZY, Summers RM. Tienet: text-image embedding network for common thorax disease classification and reporting in chest X-rays. Proc Cvpr IEEE 2018:9049–58. https://doi.org/10.1109/Cvpr.2018.00943.

[35] Nakaura T, Higaki T, Awai K, Ikeda O, Yamashita Y. A primer for understanding radiology articles about machine learning and deep learning. Diagnost Intervent Imag 2020;101(12):765–70. https://doi.org/10.1016/j.diii.2020.10.001.

[36] Chan H-P, Hadjiiski LM, Samala RK. Computer-aided diagnosis in the era of deep learning. Med Phys 2020;47(5):e218–27. https://doi.org/10.1002/mp.13764.

[37] Varges S, Bieler H, Stede M, Faulstich LC, Irsig K, Atalla M. Semscribe: natural Language Generation for medical reports. Istanbul, Turkey: European Language Resources Association (ELRA); 2012 May.

[38] Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Briefings Bioinf 2017;19(6):1236–46. https://doi.org/10.1093/bib/bbx044.

[39] Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, Cui C, Corrado G, Thrun S, Dean J. A guide to deep learning in healthcare. Nat Med 2019;25(1):24–9. https://doi.org/10.1038/s41591-018-0316-z.

[40] Zhou SK, Greenspan H, Davatzikos C, Duncan JS, Ginneken BV, Madabhushi A, Prince JL, Rueckert D, Summers RM. A review of deep learning in medical imaging: imaging traits, technology trends, case studies with progress highlights, and future promises. Proc IEEE 2021;109(5):820–38. https://doi.org/10.1109/JPROC.2021.3054390.

[41] Esteva A, Chou K, Yeung S, Naik N, Madani A, Mottaghi A, Liu Y, Topol E, Dean J, Socher R. Deep learning-enabled medical computer vision. npj Dig Med 2021;4(1):5. https://doi.org/10.1038/s41746-020-00376-2.

[42] Sorin V, Barash Y, Konen E, Klang E. Deep learning for Natural Language Processing in radiology—fundamentals and a systematic review. J Am Coll Radiol 2020;17(5):639–48. https://doi.org/10.1016/j.jacr.2019.12.026.

[43] Chan H-P, Samala RK, Hadjiiski LM, Zhou C. Deep learning in medical image analysis. In: Lee G, Fujita H, editors. Deep learning in medical image analysis :

challenges and applications. Cham: Springer International Publishing; 2020. p. 3–21.
[44] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Se Brennan R Chou, Glanville J, Grimshaw JM, Hróbjartsson A, Lalu MM, Li T, Loder EW, Mayo-Wilson E, McDonald S, McGuinness LA, Stewart LA, Thomas J, Tricco AC, Welch VA, Whiting P, Moher D. The prisma 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. https://doi.org/10.1136/bmj.n71.
[45] Clarivate. Web of science. No date, https://clarivate.com/webofsciencegroup/solutions/web-of-science/. [Accessed 12 April 2022].
[46] National Center for Biotechnology Information. National library of medicine. Pubmed. No date, https://pubmed.ncbi.nlm.nih.gov/. [Accessed 12 April 2022].
[47] Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960; 20(1):37–46. https://doi.org/10.1177/001316446002000104.
[48] Yang Y, Cao J, Wen Y, Zhang P. Table to text generation with accurate content copying. Sci Rep 2021;11(1):22750. https://doi.org/10.1038/s41598-021-00813-6. 50.
[49] Demner-Fushman D, Kohli MD, Rosenman MB, Shooshan SE, Rodriguez L, Antani S, Thoma GR, McDonald CJ. Preparing a collection of radiology examinations for distribution and retrieval. J Am Med Inf Assoc 2016;23(2): 304–10. https://doi.org/10.1093/jamia/ocv080 [published Online First: 20150701].
[50] Spinks G, Moens MF. Justifying diagnosis decisions by deep neural networks. J Biomed Inf 2019;96:103248. https://doi.org/10.1016/j.jbi.2019.103248 [published Online First: 20190706].
[51] Xiong YX, Du B, Yan PK. Reinforced transformer for medical image captioning. Machine learning in medical imaging (mlmi 2019), vol. 11861; 2019. p. 673–80. https://doi.org/10.1007/978-3-030-32692-0_77.
[52] Huang X, Yan FQ, Xu W, Li MZ. Multi-attention and incorporating background information model for chest X-ray image report generation. IEEE Access 2019;7: 154808–17. https://doi.org/10.1109/access.2019.2947134.
[53] Gasimova A. Automated enriched medical concept generation for chest X-ray images. In: 2nd international workshop on interpretability of machine intelligence in medical image computing (IMIMIC)/9th international workshop on multimodal learning for clinical decision support (ML-CDS). Shenzhen, PEOPLES R CHINA: Springer International Publishing Ag; 2019 Oct 17.
[54] Yuan JB, Liao HF, Luo R, Luo JB. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In: 10th international workshop on machine learning in medical imaging (MLMI)/22nd international conference on medical image computing and computer-assisted intervention (MICCAI). Shenzhen, PEOPLES R CHINA: Springer International Publishing Ag; 2019 Oct 13-17.
[55] Singh S, Karimi S, Ho-Shon K, Hamey L. From chest X-rays to radiology reports: a multimodal machine learning approach. In: APRS international conference on digital image computing - techniques and applications (DICTA). Perth, AUSTRALIA: Ieee; 2019 Dec 02-04.
[56] Zhang YX, Wang XS, Xu ZY, Yu QH, Yuille A, Xu DG. When radiology report generation meets knowledge graph. 34th AAAI conference on artificial intelligence. In: 32nd innovative applications of artificial intelligence conference/ 10th AAAI symposium on educational advances in artificial intelligence. New York, NY: Assoc Advancement Artificial Intelligence; 2020 Feb 07-12.
[57] 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.
[58] Xue Y, Huang XL. Improved disease classification in chest X-rays with transferred features from report generation. In: 26th biennial international conference on information processing in medical imaging (IPMI). Hong Kong, HONG KONG: Springer International Publishing Ag; 2019 Jun 02-07. Hong Kong Univ Sci & Technol.
[59] Liu G, Hsu T-MH, McDermott M, Boag W, Weng W-H, Szolovits P, Ghassemi M. Clinically accurate chest X-ray report generation. In: Machine learning for healthcare conference. PMLR; 2019.
[60] Yin C, Qian B, Wei J, Li X, Zhang X, Li Y, Zheng Q. Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network. In: 2019 IEEE international conference on data mining (ICDM); 2019 8-11 Nov. 2019.
[61] Jing B, Wang Z, Xing E. Show, describe and conclude: on exploiting the structure information of chest X-ray reports. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019.
[62] Alfarghaly O, Khaled R, Elkorany A, Helal M, Fahmy A. Automated radiology report generation using conditioned transformers. Inform Med Unlocked 2021; 24:100557. https://doi.org/10.1016/j.imu.2021.100557.
[63] Nooralahzadeh F, Gonzalez NP, Frauenfelder T, Fujimoto K, Krauthammer M. Progressive transformer-based generation of radiology reports. 2021. arXiv preprint arXiv:2102.09777.
[64] Harzig P, Chen Y-Y, Chen F, Lienhart R. Addressing data bias problems for chest X-ray image report generation. 2019. arXiv preprint arXiv:1908.02123.
[65] Li X, Cao R, Zhu D. Vispi: automatic visual perception and interpretation of chest X-rays. 2019. arXiv preprint arXiv:1906.05190.
[66] Johnson AEW, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng C-y, Mark RG, Horng S. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data 2019;6(1):317. https://doi.org/10.1038/s41597-019-0322-0.
[67] Johnson AEW, Pollard TJ, Greenbaum NR, Lungren MP, Deng C-y, Peng Y, Lu Z, Mark RG, Berkowitz SJ, Horng S. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. 2019. https://ui.adsabs.harvard.edu/abs/2019arXiv190107042J. [Accessed 1 January 2019].
[68] Rodin I, Fedulova I, Shelmanov A, Dylov DV. Multitask and multimodal neural network model for interpretable analysis of X-ray images. In: IEEE international conference on bioinformatics and biomedicine (BIBM). San Diego, CA: Ieee; 2019 Nov 18-21.
[69] Lovelace J, Mortazavi B. Learning to generate clinically coherent chest X-ray reports. In: Proceedings of the 2020 conference on empirical methods in Natural Language Processing: findings; 2020.
[70] Moreira IC, Amaral I, Domingues I, Cardoso A, Cardoso MJ, Inbreast JS Cardoso. Toward a full-field digital mammographic database. Acad Radiol 2012;19(2): 236–48. https://doi.org/10.1016/j.acra.2011.09.014 [published Online First: 20111110].
[71] Sun L, Wang W, Li J, Lin J. Study on medical image report generation based on improved encoding-decoding method. In: Huang DS, Bevilacqua V, Premaratne P, editors. 15th international conference on intelligent computing, ICIC 2019. Springer Verlag; 2019. p. 686–96.
[72] Yan K, Wang X, Lu L, Summers R. Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. J Med Imag 2018;5(3):036501.
[73] Yan K, Peng YF, Lu ZY, Summers RM. Fine-grained lesion annotation in ct images with knowledge mined from radiology reports. In: 16th IEEE international symposium on biomedical imaging (ISBI); 2019 Apr 08-11. Venice, ITALY.
[74] Marvasti NB, Garcia MdMR, Üsküdarli S, Montes JFA, Acar B. Overview of the imageclef 2015 liver ct annotation task. CLEF (Working Notes); 2015.
[75] Loveymi S, Dezfoulian MH, Mansoorizadeh M. Automatic generation of structured radiology reports for volumetric computed tomography images using question-specific deep feature extraction and learning. J Med Signals Sens 2021; 11(3):194–207. https://doi.org/10.4103/jmss.JMSS_21_20 [published Online First: 20210721].
[76] National Library of Medicine. Medical subject headings. 2021. https://www.nlm.nih.gov/mesh/meshhome.html. [Accessed 12 April 2022].
[77] Wang XS, Peng YF, Lu L, Lu ZY, 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: 30th ieee conference on computer vision and pattern recognition (cvpr 2017); 2017. p. 3462–71. https://doi.org/10.1109/Cvpr.2017.369.
[78] Irvin J, Rajpurkar P, Ko M, Yu YF, Ciurea-Ilcus S, Chute C, Marklund H, Haghgoo B, Ball R, Shpanskaya K, Seekins J, Mong DA, Halabi SS, Sandberg JK, Jones R, Larson DB, Langlotz CP, Patel BN, Lungren MP, Ng AY. Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. 33rd aaai conference on artificial intelligence/31st innovative applications of artificial intelligence conference/9th aaai symposium on educational advances in artificial intelligence. Honolulu, HI: Assoc Advancement Artificial Intelligence; 2019 Jan .
[79] Han ZY, Wei BZ, Leung S, Chung J, Li S. Towards automatic report generation in spine radiology using weakly supervised framework. In: 21st international conference on medical image computing and computer-assisted intervention (MICCAI)/8th eurographics workshop on visual computing for biology and medicine (VCBM)/international workshop on computational diffusion MRI (CDMRI). Granada, SPAIN: Springer International Publishing Ag; 2018 Sep 16-21.
[80] Ku Lewandrowski, Muraleedharan N, Eddy SA, Sobti V, Reece BD, Ramirez Leon JF, Shah S. Feasibility of deep learning algorithms for reporting in routine spine magnetic resonance imaging. Internet J Spine Surg 2020;14(s3):S86–97. https://doi.org/10.14444/7131.
[81] Zeng XH, Wen L, Liu BG, Qi XJ. Deep learning for ultrasound image caption generation based on object detection. Neurocomputing 2020;392:132–41.
[82] Zeng XH, Liu BG, Zhou M. Understanding and generating ultrasound image description. J Comput Sci Tech-Ch 2018;33(5):1086–100. https://doi.org/10.1007/s11390-018-1874-8.
[83] Alsharid M, Sharma H, Drukker L, Chatelain P, Papageorghiou AT, Noble JA. Captioning ultrasound images automatically. Med Image Comput Comput Assist Interv 2019;22:338–46. https://doi.org/10.1007/978-3-030-32251-9_37 [published Online First: 20191010].
[84] Gale W, Oakden-Rayner L, Carneiro G, Bradley AP, Palmer LJ. Detecting hip fractures with radiologist-level performance using deep neural networks. 2017. https://ui.adsabs.harvard.edu/abs/2017arXiv171106504G. [Accessed 1 November 2017].
[85] Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. Adv Neural Inf Process Syst 2014;27.
[86] 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.
[87] Ku Lewandrowski, Muraleedharan N, Eddy SA, Sobti V, Reece BD, Ramirez Leon JF, Shah S. Reliability analysis of deep learning algorithms for reporting of routine lumbar mri scans. Internet J Spine Surg 2020;14(s3):S98–107. https://doi.org/10.14444/7132 [published Online First: 20201029].
[88] Gale W, Oakden-Rayner L, Carneiro G, Palmer LJ, Bradley AP. Producing radiologist-quality reports for interpretable deep learning. In: 16th IEEE international symposium on biomedical imaging (ISBI). Venice, ITALY: Ieee; 2019 Apr 08-11.
[89] 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.
[90] Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y. Learning phrase representations using rnn encoder–decoder for statistical machine translation. Doha, Qatar: Association for Computational Linguistics; 2014 oct.

[91] Hermans M, Schrauwen B. Training and analysing deep recurrent neural networks. Adv Neural Inf Process Syst 2013;26.
[92] Pascanu R, Gulcehre C, Cho K, Bengio Y. How to construct deep recurrent neural networks. 2013. arXiv preprint arXiv:1312.6026.
[93] Krause J, Johnson J, Krishna R, Fei-Fei L. A hierarchical approach for generating descriptive image paragraphs. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR); 2017 21-26 July 2017.
[94] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Adv Neural Inf Process Syst 2017.
[95] Goodfellow I, J Pouget-Abadie M Mirza, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. Adv Neural Inf Process Syst 2014;27.
[96] Yao T, Pan Y, Li Y, Mei T. Exploring visual relationship for image captioning. In: Computer vision – eccv 2018; 2018. Cham: Springer International Publishing; 2018.
[97] Wang D, Beck D, Cohn T. On the role of scene graphs in image captioning. Hong Kong, China: Association for Computational Linguistics; 2019 November.
[98] Chang X, Ren P, Xu P, Li Z, Chen X, Hauptmann A. A comprehensive survey of scene graphs: generation and application. Ieee T Pattern Anal 2023;45(1):1–26. https://doi.org/10.1109/TPAMI.2021.3137605.
[99] Rs Sutton AG Barto. Reinforcement learning: an introduction. MIT press; 2018.
[100] Williams RJ. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Mach Learn 1992;8(3):229–56. https://doi.org/10.1007/BF00992696.
[101] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. 2014. arXiv preprint arXiv:1409.0473.
[102] Xu K, Ba JL, Kiros R, Cho K, Courville A, Salakhutdinov R, Zemel RS, Bengio Y. Show, attend and tell: neural image caption generation with visual attention. In: 32nd international conference on machine learning, ICML 2015. International Machine Learning Society (IMLS); 2015.
[103] Ronneberger O, Fischer P, U-Net T Brox. Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer; 2015.
[104] Zhao J, Kim Y, Zhang K, Rush A, LeCun Y. Adversarially regularized Autoencoders. In: Jennifer D, Andreas K, editors. Proceedings of the 35th international conference on machine learning. Proceedings of machine learning research. PMLR; 2018. p. 5902–11.
[105] Lu J-T, Pedemonte S, Bizzo B, Doyle S, Andriole KP, Michalski MH, Gonzalez RG, Pomerantz SR. Deep spine: automated lumbar vertebral segmentation, disc-level designation, and spinal stenosis grading using deep learning. In: Machine learning for healthcare conference. PMLR; 2018.
[106] Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: a method for automatic evaluation of machine translation. 40th annual meeting of the association for computational linguistics. Proceed Conf 2002:311–8. https://doi.org/10.3115/1073083.1073135.
[107] Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017.
[108] Lin C-Y. Rouge: a package for automatic evaluation of summaries. Barcelona, Spain: Association for Computational Linguistics; 2004 jul.
[109] Denkowski M, Lavie A. Meteor universal: language specific translation evaluation for any target language. Baltimore, Maryland, USA: Association for Computational Linguistics; 2014 jun.
[110] Likert R. A technique for the measurement of attitudes. Archiv psychol 1932;140:1–55.
[111] Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull 1971;76(5):378–82. https://doi.org/10.1037/h0031619.
[112] Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951;16(3):297–334. https://doi.org/10.1007/BF02310555.
[113] Rennie SJ, Marcheret E, Mroueh Y, Ross J, Goel V. Self-critical sequence training for image captioning. In: 30th ieee conference on computer vision and pattern recognition (cvpr 2017); 2017. p. 1179–95. https://doi.org/10.1109/Cvpr.2017.131.
[114] Donahue J, Hendricks LA, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, Darrell T. Long-term recurrent convolutional networks for visual recognition and description. In: 2015 ieee conference on computer vision and pattern recognition (cvpr); 2015. p. 2625–34.
[115] Lu JS, Xiong CM, Parikh D, Socher R. Knowing when to look: adaptive attention via a visual sentinel for image captioning. In: 30th ieee conference on computer vision and pattern recognition (cvpr 2017); 2017. p. 3242–50. https://doi.org/10.1109/Cvpr.2017.345.
[116] Spasic I, Nenadic G. Clinical text data in machine learning: systematic review. JMIR Med Inf 2020;8(3):e17984. https://doi.org/10.2196/17984.
[117] Giménez J, Màrquez L. Linguistic features for automatic evaluation of heterogenous Mt systems. Prague, Czech Republic: Association for Computational Linguistics; 2007 June.
[118] Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, Liu T-Y. Biogpt: generative pre-trained transformer for biomedical text generation and mining. Briefings Bioinf 2022;23(6):bbac409. https://doi.org/10.1093/bib/bbac409.
[119] Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shya P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D. Language Models are few-shot learners. Adv Neural Inf Process Syst 2020;33:1877–901.
[120] European Society of Radiology (ESR). Esr paper on structured reporting in radiology. Insights Imag 2018;9(1):1–7. https://doi.org/10.1007/s13244-017-0588-8.
[121] Jm Nobel K van Geel, Robben SGF. Structured reporting in radiology: a systematic review to explore its potential. Eur Radiol 2022;32(4):2837–54. https://doi.org/10.1007/s00330-021-08327-5.
[122] Anderson P, Fernando B, Johnson M, Gould S. Spice: semantic propositional image caption evaluation. In: Computer vision–ECCV 2016: 14th European conference, amsterdam, The Netherlands, october 11-14, 2016, proceedings, Part V 14. Springer; 2016.
[123] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017.
[124] Zhang Y, Chen Q, Yang Z, Lin H, Lu Z. Biowordvec, improving biomedical word embeddings with subword information and mesh. Sci Data 2019;6(1):1–9.
[125] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. arXiv preprint arXiv:1409.1556.
[126] 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.
[127] Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. Imagenet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition; 2009 20-25 june; 2009.
[128] Jégou S, Drozdzal M, Vazquez D, Romero A, Bengio Y. The one hundred layers tiramisu: fully convolutional densenets for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops; 2017.
[129] Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Mobilenets H Adam. Efficient convolutional neural networks for mobile vision applications. 2017. arXiv preprint arXiv:1704.04861.
[130] LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD. Backpropagation applied to handwritten zip code recognition. Neural Comput 1989;1(4):541–51. https://doi.org/10.1162/neco.1989.1.4.541.
[131] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2015.
[132] Pyysalo S, Ginter F, Moen H, Salakoski T, Ananiadou S. Distributional semantics resources for biomedical text processing. Proceed LBM 2013:39–44.
[133] Rehurek R. The gensim Word2vec implementation. No date, https://radimrehurek.com/gensim/models/word2vec.html. [Accessed 12 April 2022].
[134] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: visual explanations from deep networks via gradient-based localization. In: 2017 IEEE international conference on computer vision (ICCV); 2017 22-29 oct; 2017.
[135] Google Code. The google code Word2vec implementation. 2013. . [Accessed 12 April 2022].
[136] Pennington J, Socher R, Manning CD. Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing. EMNLP); 2014.
[137] Ren SQ, He KM, Girshick R, Sun J. Faster R-cnn: towards real-time object detection with region proposal networks. Ieee T Pattern Anal 2017;39(6):1137–49. https://doi.org/10.1109/Tpami.2016.2577031.
[138] Ordonez V, Kulkarni G, Berg TL. Im2text: describing images using 1 million captioned photographs. In: Proceedings of the 24th international conference on neural information processing systems. Granada, Spain: Curran Associates Inc.; 2011. p. 1143–51.
[139] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016.
[140] Zhang Y, Ding DY, Qian T, Manning CD, Langlotz CP. Learning to summarize radiology findings. EMNLP 2018 2018:204.
[141] Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional lstm and other neural network architectures. Neural Network 2005;18(5):602–10. https://doi.org/10.1016/j.neunet.2005.06.042.
[142] Conneau A, Kiela D, Schwenk H, Barrault L, Bordes A. Supervised learning of universal sentence representations from Natural Language inference data. 2017. arXiv preprint arXiv:1705.02364.
[143] Zhang Y, Ding DY, Qian T, Manning CD, Langlotz CP. Learning to summarize radiology findings. Brussels, Belgium: Association for Computational Linguistics; 2018 oct.
[144] Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. Adv Neural Inf Process Syst 2013.
[145] Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, Ding D, Bagul A, Langlotz C, Shpanskaya K, Lungren MP, Ng AY. Chexnet: radiologist-level pneumonia detection on chest X-rays with deep learning. 2017. arXiv preprint arXiv:1711.05225.
[146] McDonald R, Brokos G, Androutsopoulos I. Deep relevance ranking using enhanced document-query interactions. Brussels, Belgium: Association for Computational Linguistics; 2018.
[147] HuggingFace. Distilgpt2. No date, https://huggingface.co/distilgpt2. [Accessed 15 January 2022].
[148] Gokaslan A, Cohen V. Openwebtext corpus. 2019. http://Skylion007.github.io/OpenWebTextCorpus. [Accessed 28 December 2021].
[149] Cornia M, Stefanini M, Baraldi L, Cucchiara R. Meshed-memory transformer for image captioning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2020.

[150] Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, Stoyanov V, Bart L Zettlemoyer. Denoising sequence-to-sequence pre-training for Natural Language Generation, translation, and comprehension. Online. Association for Computational Linguistics; 2020 jul.
[151] Rehurek R, Sojka P. Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks. Citeseer; 2010.
[152] Zhang H, Xu T, Li H, Zhang S, Wang X, Huang X, Metaxas DN. Stackgan: text to photo-realistic image synthesis with stacked generative adversarial networks. Proceedings of the IEEE international conference on computer vision 2017.
[153] Zagoruyko S, Komodakis N. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. 2016. arXiv preprint arXiv:1612.03928.
[154] Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. 2016. arXiv preprint arXiv:1609.02907.