



Survey paper

From vision to text: A comprehensive review of natural image captioning in medical diagnosis and radiology report generation

Gabriel Reale-Nosei^a, Elvira Amador-Domínguez^{b,c,*}, Emilio Serrano^b

^a ETSI Informáticos, Universidad Politécnica de Madrid, 28660 Boadilla del Monte, Madrid, Spain

^b Ontology Engineering Group, Departamento de Inteligencia Artificial, ETSI Informáticos, Universidad Politécnica de Madrid, 28660 Boadilla del Monte, Madrid, Spain

^c Departamento de Sistemas Informáticos, ETSI Sistemas Informáticos, Universidad Politécnica de Madrid, 28031 Madrid, Spain

ARTICLE INFO

Keywords:

Medical image captioning
Natural image captioning
Diagnostic captioning
Radiology report generation
Survey
State-of-the-art review

ABSTRACT

Natural Image Captioning (NIC) is an interdisciplinary research area that lies within the intersection of Computer Vision (CV) and Natural Language Processing (NLP). Several works have been presented on the subject, ranging from the early template-based approaches to the more recent deep learning-based methods. This paper conducts a survey in the area of NIC, especially focusing on its applications for Medical Image Captioning (MIC) and Diagnostic Captioning (DC) in the field of radiology. A review of the state-of-the-art is conducted summarizing key research works in NIC and DC to provide a wide overview on the subject. These works include existing NIC and MIC models, datasets, evaluation metrics, and previous reviews in the specialized literature. The revised work is thoroughly analyzed and discussed, highlighting the limitations of existing approaches and their potential implications in real clinical practice. Similarly, future potential research lines are outlined on the basis of the detected limitations.

Contents

1. Introduction	2
2. Research method	3
2.1. Keyword selection	3
2.2. Resource retrieval	3
2.3. Resource selection	3
2.4. Resource analysis	4
3. Natural image captioning	4
3.1. Surveys	5
3.2. Early approaches	5
3.2.1. Retrieval-based methods	5
3.2.2. Template-based methods	5
3.3. Deep learning based approaches	6
3.3.1. Encoder-Decoder architectures	6
3.3.2. Compositional architectures	6
3.3.3. Attention-based architectures	7
3.3.4. Dense image captioning	7
4. Medical image captioning and diagnosis	7
4.1. Datasets for radiological report generation	8
4.2. Evaluation methods	9
4.2.1. Natural language generation evaluation methods	9
4.2.2. Clinical efficacy evaluation methods	10
4.3. Medical report generation methods	10

* Corresponding author at: Ontology Engineering Group, Departamento de Inteligencia Artificial, ETSI Informáticos, Universidad Politécnica de Madrid, 28660 Boadilla del Monte, Madrid, Spain.

E-mail addresses: gabriel.reale.nosei@alumnos.upm.es (G. Reale-Nosei), elvira.amador@upm.es (E. Amador-Domínguez), emilio.serrano@upm.es (E. Serrano).

<https://doi.org/10.1016/j.media.2024.103264>

Received 9 August 2023; Received in revised form 25 April 2024; Accepted 1 July 2024

Available online 8 July 2024

1361-8415/© 2024 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

4.3.1.	Retrieval-based methods	10
4.3.2.	Template-based methods	10
4.3.3.	Generative methods	11
4.3.4.	Hybrid methods	12
5.	Discussion	12
5.1.	Image captioning and medical report generation	12
5.2.	Medical image captioning	13
5.2.1.	Available datasets	13
5.2.2.	Evaluation methods	13
5.2.3.	Diagnostic captioning methods	14
5.3.	Limitations	15
5.4.	Implications in real clinic practices	16
6.	Conclusions and future work	16
	CRedit authorship contribution statement	17
	Declaration of competing interest	17
	Data availability	17
	Acknowledgments	17
	References	17

1. Introduction

Image captioning is a field that combines Computer Vision (CV) and Natural Language Processing (NLP) techniques to generate natural language descriptions of images. Image captioning has numerous applications that span across various domains, such as industry or, as studied in this paper, medicine. In the context of medical applications, this practice extends to Medical Image Captioning (MIC), which specifically involves providing textual descriptions of the contents of medical images. Diagnostic Captioning (DC) is a subset of MIC, going a step further to provide a diagnostic interpretation of the image. This involves not only describing what is seen but also conveying a judgment about what those observations mean in terms of a diagnosis. Although MIC is a broader term that encompasses tasks beyond formal diagnoses, MIC and DC are often used interchangeably in the literature, mainly because describing or listing healthy organs is often not as clinically relevant (see Fig. 1).

Medical image captioning is particularly notable in the radiology field. One of the prime tasks in this is automatic report generation, namely, generating medical reports automatically from a set of images. Medical reports offer diagnostic interpretations, patient history, comparisons with previous studies, and recommendations for further tests or treatments. Thus, automatically generating medical reports encompasses, but is not necessarily limited to, Diagnostic Captioning (DC). Automating the generation of reports offers several advantages, one of which is its improved repeatability. With image captioning, healthcare professionals can achieve consistent and standardized reporting in diverse cases, promoting accuracy and reliability. Furthermore, by employing a structured report in a human-readable format, image captioning offers clear and comprehensive reports that facilitate effective communication among healthcare professionals. Therefore, physicians can rely on this technology to serve as a second-hand source of diagnoses or findings.

The image captioning task involves tackling two fundamental questions: visual understanding and linguistic processing, as it represents the convergence of CV and NLP. In the field of natural image captioning, different approaches have been proposed, including retrieval-based, template-based, neural network-based, or attention-based, among others. These methods take advantage of the information extracted from the image to generate captions that accurately represent the content of the image. For Medical Image Captioning (MIC), there are specific approaches, tailored to the unique features and requirements of radiology reports. In particular, this paper focuses on methods related to Diagnostic Captioning (DC) within medical radiology reports. Unlike MIC, DC has an influence on the diagnostic interpretation of medical images. When referring to automatic generation of medical

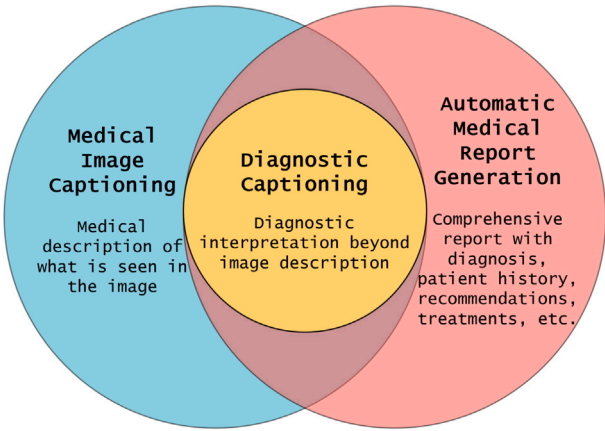


Fig. 1. Visual representation of the definition of Medical Image Captioning (MIC), Diagnostic Captioning (DC), and automatic medical report generation.

reports, the discussion within this article goes beyond a diagnostic statement to include image comparisons of previous or multiple studies, patient history, and other pertinent aspects. The survey presented herein comprehensively explores techniques for both Natural Image Captioning (NIC) and distinct strategies for Diagnostic Captioning (DC). Discussions on more advanced medical report generation methods are also provided.

Radiology reports play a critical role in conveying diagnosis findings and guiding patient care. It summarizes the findings and interpretations of radiological imaging studies performed on a patient after a medical visit. It serves as a vital communication tool between radiologists and referring physicians, providing essential information for diagnosis, treatment planning, and patient management. A typical radiology report consists of several sections, each containing specific information. Fig. 2 shows an example of a radiology report, extracted from the MIMIC-CXR MIC data set (Johnson et al., 2019b). Each section of the report plays a vital role in providing accurate and comprehensive information on efficient documentation and communication.

The main goal of this paper is to present an overview of the current state of the art in NIC, focusing particularly on DC and its applications in the field of radiology. By reviewing the existing literature and research, this work offers information on the advances made in this field, identifies challenges and limitations, and explores potential directions for improvement. Moreover, this work attempts to identify key seminal papers for NIC that served as a basis for DC. It aims to establish, as well, the common ground and differences between both approaches, offering alternatives and future research directions to further benefit

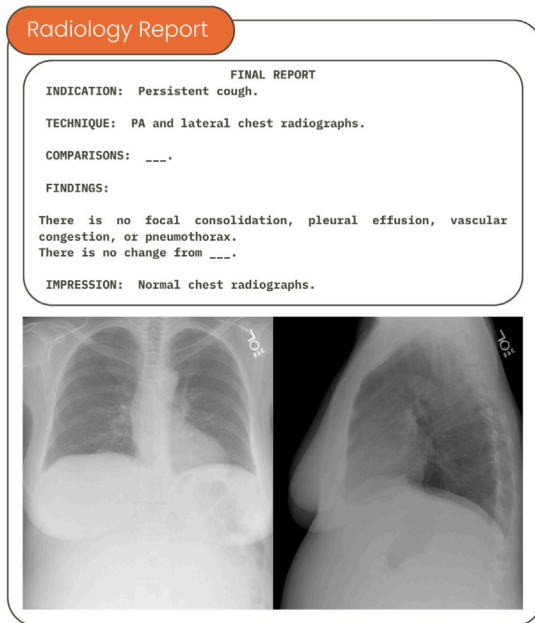


Fig. 2. Example of a radiology report, as formatted in the MIMIC-CXR dataset (Johnson et al., 2019b).

DC towards its final goal of automatically generating complete medical reports.

The paper is structured as follows. Section 2 outlines the research method followed to extract, filter, and analyze existing relevant resources from the literature. Section 3 presents and discusses NIC related work, while Section 4 reviews and presents those works related to Medical Image Captioning. Section 5 discusses the main aspects extracted after the analysis of the recovered resources, and Section 6 draws conclusions and potential future research lines.

2. Research method

The conducted survey first focuses on Nature Image Captioning (NIC) as a whole and further explores its applications in the context of medical image, especially in the context of radiology. The survey is conducted following a four-step approach. First, search keywords are specified to target the search and retrieve relevant resources for the study. These keywords are then searched in the main scientific repositories: Web of Science, Scopus, Arxiv, and Google Scholar. This leads to a set of unrevised potential resources, both NIC and MIC related. As explained above, the terms MIC and DC are generally used interchangeably in the literature. In this section, MIC is preferred to integrate these medical-related methods into a broader term and avoid missing out on relevant publications with either naming convention. The inclusion and exclusion criteria are then applied to obtain the final set of related resources, which are further analyzed and summarized in this paper. Fig. 3 shows the aforementioned stages.

2.1. Keyword selection

The selection of appropriate keywords is essential to carry out a comprehensive and targeted search, ensuring the inclusion of relevant and related works while minimizing the risk of overlooking pertinent literature. A systematic approach is followed to identify the appropriate keywords. First, the research topic was explored to determine the general terms related to it, drawing on existing literature and domain expertise. This initial set of terms was further refined to include synonyms, related terms, and commonly used variations, as well as the recurring keywords featured in the retrieved papers. The final selection

Table 1

Selected query keywords for NIC and MIC related resources.

NIC keywords	MIC keywords
"image captioning"	"medical"
"visual captioning"	"radiology report"
"image-to-text generation"	"diagnostic report"
"caption generation"	
"visual content description"	

of keywords poses a balance in which the number of relevant articles retrieved is maximized while minimizing the amount of noise that exists in the search results.

As described in Fig. 3, keywords for searching both NIC-related and MIC-related terms are specified in this first stage. Table 1 outlines the keywords associated with each topic. In the NIC search (Listing 1), all terms are treated disjunctively, and therefore papers related to this topic can be retrieved using the query:

Listing 1: "Search query for NIC resources"

```
"image captioning" OR "visual
captioning" OR "image-to-text
generation" OR "caption generation"
OR "visual content description"
```

Nonetheless, MIC resources require from a more targeted search. As MIC is a subfield of NIC, the same keywords are used in addition to the specific MIC keywords depicted in Table 1. Listing 2 outlines the query used to retrieve the resources specifically related to MIC.

Listing 2: "Search query for MIC resources"

```
("medical" OR "radiology report" OR "
diagnostic report") AND ("image
captioning" OR "visual captioning"
OR "image-to-text generation" OR "
caption generation" OR "visual
content description")
```

2.2. Resource retrieval

Once the keywords and search queries have been specified, the resources can be searched and recovered. As described in the resource retrieval stage in Fig. 3, four databases were considered for the search. Web of Science and SCOPUS are two of the most widely used scientific databases, containing publications from different indexed sources. Being two of the primary bibliometric sources, resources are first retrieved from them. Non-indexed databases are also considered to expand the scope of the search. For this purpose, both Arxiv and Google Scholar are also selected as data sources.

A diverse range of resources, covering different aspects of MIC and NIC, are considered to provide a wide and comprehensive view on the topic. Therefore, articles focusing on NIC architectures, novel approaches, evaluation metrics, datasets, and applications within the medical domain are considered for discussion. Survey papers are also considered, as they provide a comprehensive understanding of the research landscape in both considered topics. The information contained in these sources then serves as a baseline for further discussion of the identified resources.

2.3. Resource selection

After retrieval, a set of potential resources, both related to the NIC and MIC, is obtained. This initial raw set has to be then filtered to identify the final set of resources, which will be considered for analysis. A set of inclusion and exclusion criteria is formulated for this

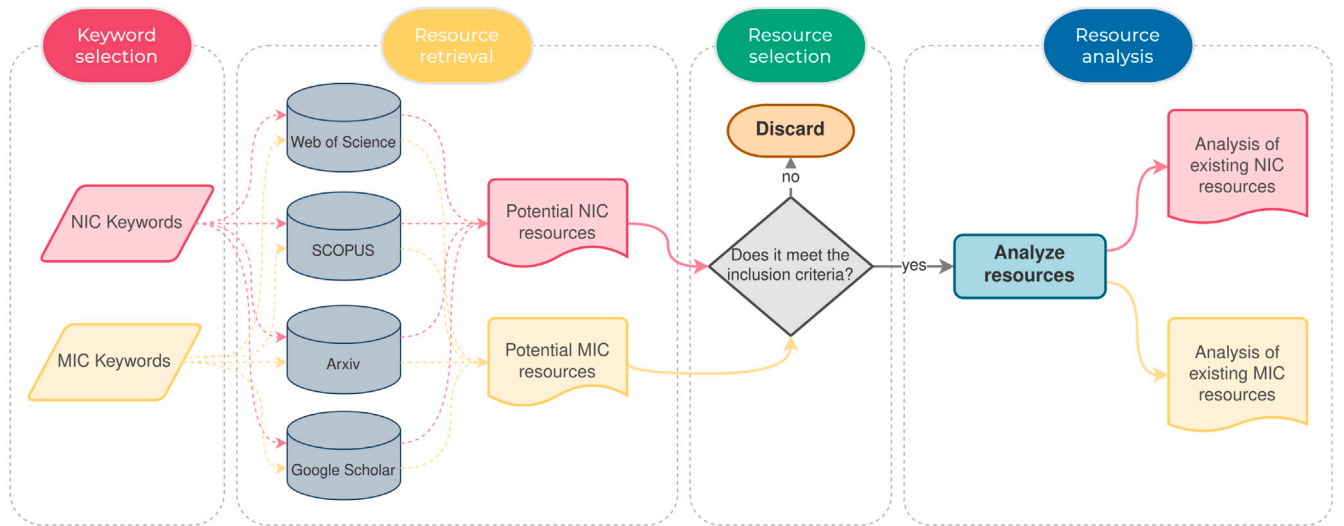


Fig. 3. Overview on the methodology followed to conduct the survey.

Table 2
Inclusion and exclusion rules.

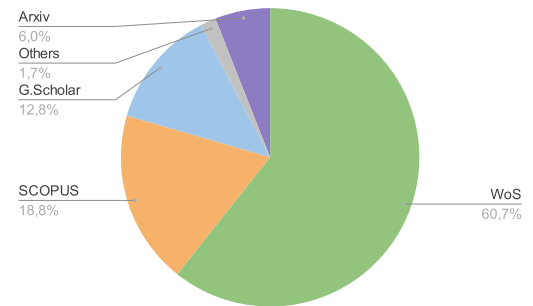
Inclusion	Exclusion
Published after 2010	Published before 2010
Published in a reputable journal or conference	Published in nonindexed journals and nonrelevant conferences
Written in English	
High impact factor and citation count	

task. Table 2 contains the inclusion and exclusion criteria applied to select the final resources to be further analyzed. A final number of 114 resources are selected from the data sets considered, distributed as reported in Fig. 4(a). As seen in the figure, most of the resources are extracted from the Web of Science database. This is directly related to the inclusion criteria, as most well-established and indexed journals are referred to in the Web of Science database. Regarding the type of paper, 38 of the resources recovered are journal articles, while 68 are conference papers. The remaining 12 resources, which are all extracted from Arxiv or Google Scholar, are not published in journals or conferences. Most of these resources are related to MIC-related data sets that fall within the scope of this survey. Therefore, despite meeting the second exclusion criterion, they are still considered.

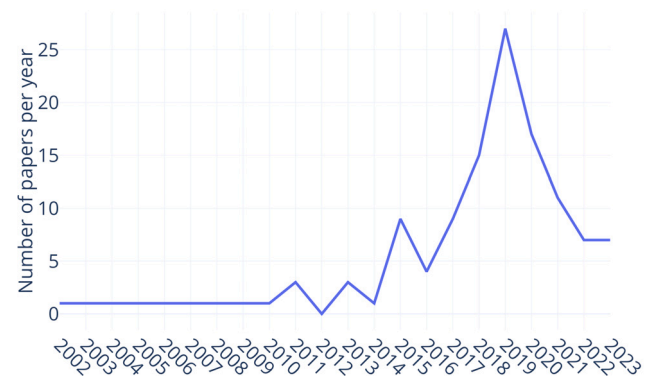
Fig. 4(b) shows the distribution of the selected resources per year. As depicted in the figure, most of the studied resources can be considered as recent, as they are dated after 2018 and, therefore, are less than five years old. It is worth noting that there are three works that date before 2010, and therefore according to the proposed exclusion criteria, they should not be further considered for analysis. These three works are related to NIC metrics and, due to their notably high relevance, are treated as an exception to the first exclusion criterion.

2.4. Resource analysis

The final stage of the survey, as illustrated in Fig. 3, comprises the analysis of the selected resources. This aspect will be addressed in depth in the subsequent sections. For easier analysis, the retrieved resources are categorized according to their type. In NIC, three main types of resources can be identified: surveys, early approaches, and deep learning-based approaches. Similarly, in MIC, resources are divided into four categories: surveys, datasets, evaluation, and report generation approaches. In both NIC and MIC, most of the retrieved resources relate to proposed models and architectures. A finer-grained categorization according to the type of approach was followed to develop each model to further explore and study each resource. Fig. 5 shows the proportion of resources studies per type and topic.



(a) Database distribution



(b) Year distribution

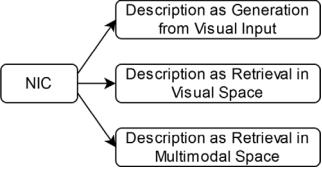
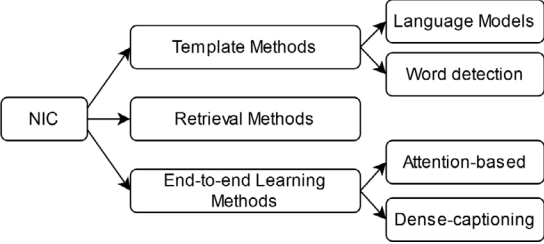
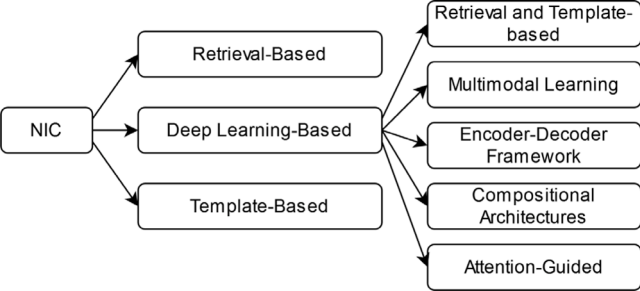
Fig. 4. Distribution of the selected resources per year and database.

3. Natural image captioning

Natural Image Captioning (NIC) is a challenging task, as it involves not only detecting and labeling concepts in an image but establishing the relationships between them. As stated in Section 2.4, the retrieved resources can be grouped into three categories: surveys, early approaches, and deep learning-based approaches.

Table 3

Summary of the reviewed NIC surveys. Fields marked with 'X' denote that the subject is covered by the survey, while '-' indicates otherwise.

Survey	Proposed taxonomy	Methods	Datasets	Evaluation
Bernardi et al. (2016)		X	X	X
Kumar and Goel (2017)	Does not propose a taxonomy	X	X	X
Liu et al. (2018)		-	X	X
Bai and An (2018)		X	X	X
Hossain et al. (2019)	Presented in Fig. 6	Only Deep-Learning based	X	X

3.1. Surveys

Many surveys exist in the literature that aim to provide an overview of the subject, including information on methods, data sets, and evaluation techniques. Some of the first attempts to provide an overview of NIC, including DL-based approaches, are the works by Bernardi et al. (2016), Bai and An (2018), and Kumar and Goel (2017). Later, the works by Liu et al. (2018), and Hossain et al. (2019) delved into more advanced DL-based methods. Table 3 summarizes the different categorizations presented in these works. Recently, with the explosion of attention-based approaches, Zohourianshahzadi and Kalita (2022) presented a review that focused on best practices for the use of deep attention-based methods for image captioning. More recently, the work by Zhang and Metaxas (2024) presents a survey on the opportunities and existing challenges regarding the application of large pre-trained models in the context of medical image captioning. Moreover, this survey addresses future lines of research in this area, such as the development of multimodality pre-trained models.

3.2. Early approaches

Prior to the explosion of deep learning, NIC relied on approaches that were not based on neural networks but on more simplistic approaches. Two main trends can be distinguished regarding early approaches: retrieval-based and template-based.

3.2.1. Retrieval-based methods

Retrieval-based methods were common in the early stages of image captioning. These methods produce a caption for an input image by retrieving one or more sentences from a predefined sentence pool. The generated description can be either a single sentence or a composition of sentences extracted from the pool. Farhadi et al. (2010) propose

a method in which a meaning space is defined by a tuple of the form (*object*, *action*, *scene*) containing information on the image. The meaning space acts as an intermediate representation space between the image and the sentence spaces. Once the input image is represented by its corresponding tuple, the semantic distance between the elements on the tuple and the existing sentences in the pool is calculated. The sentence that has the highest similarity to the input tuple is then selected as a caption.

Ordonez et al. (2011) presented *Im2Text*, a method that extracts global image descriptors from a set of web-scale collections of captioned photographs. When a new image is introduced, the images in the collection are re-ranked according to their similarity to the input. The caption of the top-match image is used as a description of the query. Other works focus on side tasks of the image description process, such as feature extraction, visual similarity, image retrieval, or caption composition (Hodosh et al., 2013; Gong et al., 2014; Sun et al., 2015). From a general perspective, retrieval-based methods are capable of generating syntactically correct captions, but cannot generate image-specific captions that are semantically correct.

3.2.2. Template-based methods

Template-based approaches are based on the use of templates, which contain a set of slots to generate captions through a syntactically and semantically constraint process. First, a set of image descriptors is detected. Then, to compose a sentence, these visual concepts are connected through sentence templates or specific grammar rules. Yang et al. (2011) proposed the use of quadruples of the form (*Nouns*, *Verb*, *Scenes*, *Prepositions*) as sentence templates and then applied a language model to the verbs, scenes, and prepositions that make up the sentence.

Similarly, Li et al. (2011) use visual descriptors to detect and extract semantic information from images, and then encode it in a triple of the format $\langle\langle adj1, obj1 \rangle, prep, \langle adj2, obj2 \rangle\rangle$. Other works follow

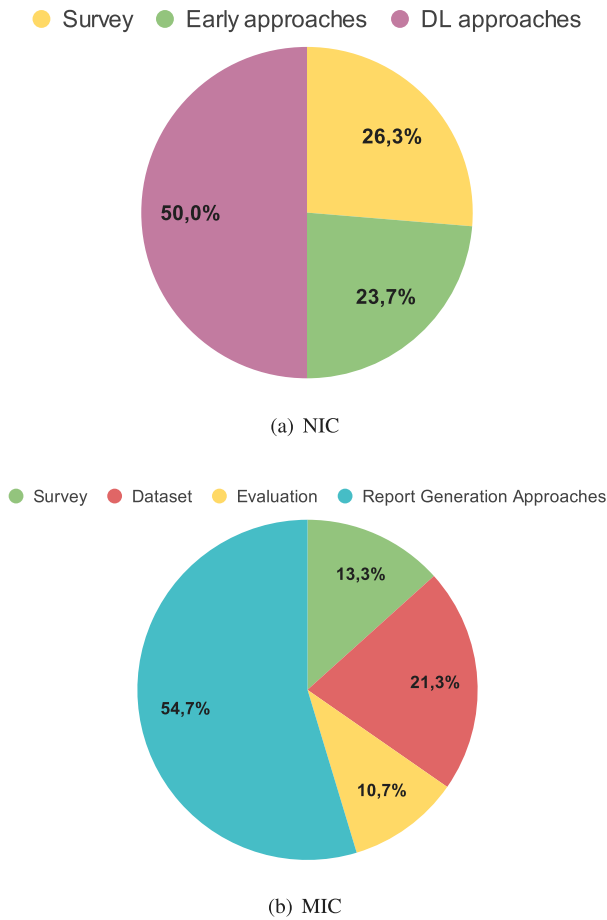


Fig. 5. Proportion of resources per type and topic.

the same approach: detecting objects in the image and composing the corresponding description by filling in a template, such as Kulkarni et al. (2013) and Ushiku et al. (2015).

Template-based methods can generate grammatically correct captions, as they are based on an initially well-defined syntactical structure. Bai and An (2018) state that the descriptions produced by template-based methods are much more semantically sound than those generated by retrieval-based methods. However, templates must be previously defined and consequently require an additional manual effort to curate and select the optimal templates for the target task. These authors also remark that template-based approaches have limited coverage, creativity, and complexity when compared to human-written captions.

3.3. Deep learning based approaches

Retrieval and template image captioning methods began to be enhanced by the progress in the area of deep learning. Instead of using hand-engineered features and shallow models as in the early approaches, further advances in computational power, data availability, alongside with the surge of deep convolutional networks enabled the use of DL models for NIC to flourish. Hossain et al. (2019) presents an exhaustive review of DL methods for image captioning in the literature. The taxonomy proposed in this paper is outlined in Fig. 6.

NICs are most often based on supervised learning-based approaches. However, since unannotated data have higher availability than labeled data, unsupervised methods have gained interest in the literature. This is the case of Generative Adversarial Networks (GANs) (Shetty et al., 2017; Dai et al., 2017), and Reinforcement Learning (RL) (Rennie

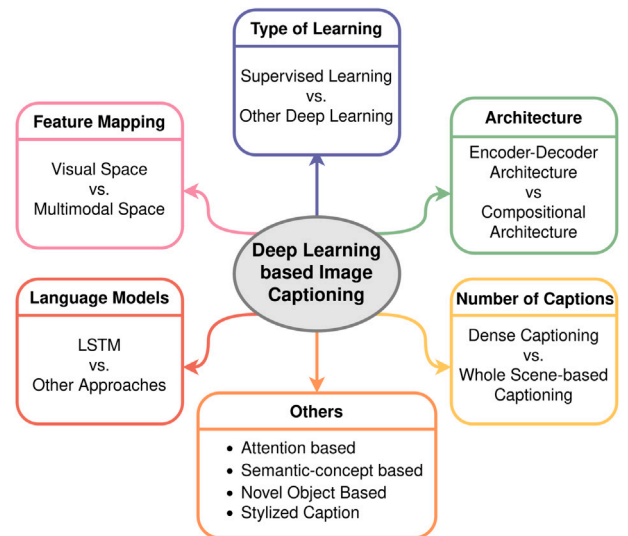


Fig. 6. A taxonomy of deep learning methods for image captioning proposed by Hossain et al. (2019).

et al., 2017; Ren et al., 2017; Zhang et al., 2017a). Although these approaches show good approximations for generating a diverse set of image captions, the best results still come from supervised learning approaches (Zohourianshahzadi and Kalita, 2022). The following subsections outline the most prominent work of the main taxonomic groups depicted in Fig. 6, with a special focus on supervised approaches.

3.3.1. Encoder–Decoder architectures

Inspired by machine translation approaches (Sutskever et al., 2014), Encoder–Decoder (ED) architectures learn features end-to-end. The *encoder*, which usually is a Convolutional Neural Network (CNN), extracts image features that are later fed into a LM to convert them into semantically and syntactically correct sentences. For example, Karpathy and Fei-Fei (2015) opt for the combination of AlexNet (Krizhevsky et al., 2012) (CNN) as the encoder, and a Recurrent Neural Network (RNN) as the decoder. Vinyals et al. (2015) propose, *Show and Tell*, which uses GoogleLeNet (Szegedy et al., 2015) as encoder, and then uses a LSTM as decoder. The revision by Hossain et al. (2019) illustrates that the selection of encoders can vary, with different CNN architectures being implemented, but LSTM is consistently used in the literature as a decoder.

3.3.2. Compositional architectures

Compositional methods group several functional building blocks that are trained independently. An input image is passed to a CNN, where the image features are obtained. These visual concepts are later fed to an LM to generate a set of candidate captions, which are then reranked by a deep multimodal similarity modal. The top pick is then selected as the image description. Similarly to ED architectures, the main challenge of compositional approaches is to find the most effective combination of building blocks. Again, Hossain et al. (2019) summarize the most relevant works that use these methods, where again LSTMs are widely used as building blocks in combination with different variants of CNN encoders. The main difference between these methods, besides not being trained in an end-to-end manner, particularly relies on the re-ranking methods and selection of the optimal output image description. For example, Fang et al. (2015) encodes the queried images using two different CNNs, and then an LM generates candidate captions that are then ranked by a linear weighting of the features of the sentence.

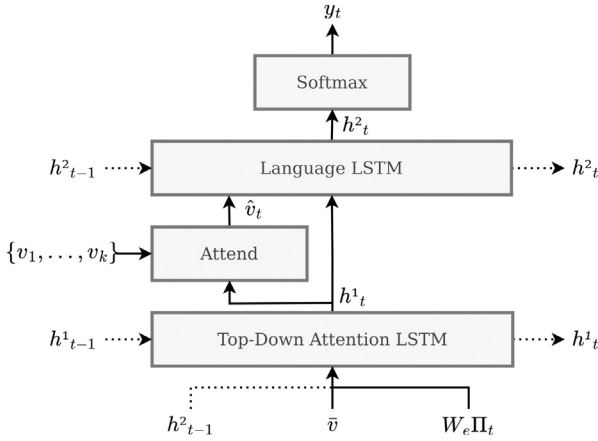


Fig. 7. Overview of the Up-Down captioning model proposed by Anderson et al. (2018). Two LSTM layers are used to selectively attend to spatial image features, \bar{v} . These features are the outputs of the Bottom-Up model, comprised by a Faster R-CNN.

3.3.3. Attention-based architectures

ED and compositional methods do not consider spatial aspects of the input image that are relevant to the output caption. Contrary to these methods, which consider the input image as a whole to generate the caption, attention-based methods can dynamically focus on pertinent regions of the input image while the output sequences are being produced. Attention-based methods have gained remarkable popularity in the context of the use of DL methods for NIC. Moreover, the work by Zohourianshahzadi and Kalita (2022) is dedicated to reviewing the state-of-the-art of attention mechanisms for NIC.

Many architectures follow the same schema as ED methods, with attention mechanisms included in the encoder, the decoder, or both. For instance, the *Show, Attend and Tell* (Xu et al., 2015) method is a variation of the *Show and Tell* method by Vinyals et al. (2015), described in Section 3.3.1. Both methods rely on a CNN to encode the input image. However, the *Show, Attend and Tell* approach incorporates attention mechanisms on the extracted visual features, which are then passed to the LSTM decoder. The inclusion of attention mechanisms greatly improved the performance of the *Show, Attend and Tell* method with respect to its non-attention counterpart.

Soft and hard attention mechanisms serve as the basis for the development of later NIC methods. As these methods rely heavily on spatial features from the whole image for attention, fine-grained details from objects and their relations are not considered. Some approaches have proposed mechanisms to address this limitation, such as semantic attention (You et al., 2016; Yao et al., 2017), spatial and channel-wise attention (Chen et al., 2017), and adaptive attention (Lu et al., 2017).

The publication of *Bottom-Up and Up-Down Attention* by Anderson et al. (2018) marked a milestone in attentive image captioning. This approach incorporates Faster R-CNN (Ren et al., 2015), which proposes several Regions of Interest, or ROIs, from the image. Then, a non-maximum suppression algorithm is applied, where regions with a detection probability lower than a given threshold are not considered for subsequent phases. This intrinsically acts as a hard, deterministic attention in a bottom-up manner, since only a limited number of ROIs are considered for visual feature extraction. Once the visual features have been extracted, they are fed into the Up-Down attention model, which is composed of two LSTMs. The first LSTM acts as a soft, top-down visual attention model on the extracted visual features. The attended values are concatenated with the hidden state attention of the first LSTM to form the input of the second LSTM. This scheme, depicted in Fig. 7, has inspired subsequent research in the field, as the quality of the generated captions considerably exceeds that of models that do not use bottom-up attention (Zohourianshahzadi and Kalita, 2022).

The Up-Down attention model achieved competitive results under all image captioning metric, this serving as a baseline for comparison with state-of-the-art models that surged afterwards. Works such as those proposed by Ke et al. (2019), Qin et al. (2019), and Wang et al. (2020) also rely on the Bottom-Up Up-Down approach. Models that employ bottom-up attention as an encoder can help the model focus on object-level details, but do not provide explicit information about the semantic or spatial relationships between them.

Alongside bottom-up attention, researchers have later incorporated Graph Convolutional Networks (GCN) (Kipf and Welling, 2017) - based methods (Yao et al., 2019; Sammani and Melas-Kyriazi, 2020), as well as MHA (Vaswani et al., 2017) - based methods (Li et al., 2019b; Cornia et al., 2020; Guo et al., 2020; Pan et al., 2020). Both methods provide the attention language model with semantic information and spatial relationships, further enhancing the quality of the captions. MHA was first introduced in the Transformer architecture by Vaswani et al. (2017), and relies on multiple scaled-dot, or self-attention heads, for neural machine translation. As reported by Zohourianshahzadi and Kalita (2022), models that employ MHA over bottom-up features and semantic information show the best results for image captioning.

3.3.4. Dense image captioning

ED and compositional methods (Sections 3.3.1 and 3.3.2) considered the images as a whole for generating captions. Attention-based approaches (Section 3.3.3) presented a series of mechanisms to focus on different parts of the image, merging the partial captions to generate a single final description of the image. However, one description of an entire scene may be subjective and not enough to provide a complete understanding of the scene. In dense captioning, captions are generated for each relevant region of the scene. *DenseCap* (Johnson et al., 2016) is one of the first models to implement this approach. In dense image caption, the typical workflow is composed of a CNN, a dense localization layer, and an LM (usually an LSTM) applied to each region feature to generate each caption.

Region-based captioning can be more objective and meticulous than global image captioning but poses a series of challenges: difficulty to recognize all target regions for each visual concept, and overlapping of multiple ROIs. Potential solutions to these issues have been proposed, such as the works of Yang et al. (2017) and Yin et al. (2019b), with the work by Shao et al. (2022) providing the most promising results. This work introduces a novel end-to-end transformer-based dense image captioning architecture that has demonstrated superiority to other state-of-the-art methods.

4. Medical image captioning and diagnosis

The literature review covered NIC and DC in great detail. However, the subsequent section primarily delves into DC due to its prevalence in the literature. Although most of the available data sets are mainly extracted from real radiological reports of medical images, most of the works published focused on simply predicting the associated captions, effectively translating into the diagnostic interpretation of medical images. Few works have explored incorporating temporal or more advanced concepts into their models to evaluate the evolution of pathologies present in such images. Consequently, the methods studied in this section are only limited to DC, as a broader concept, without delving into the particularities of radiological reports. This is the reason why the concept of DC is preferred in this section, rather than MIC. As seen in Fig. 2, a medical report is a recollection of facts or detailed statements, elaborated by an expert, that describe the relevant findings of a medical image (Monshi et al., 2020). Moreover, as already stated, the process of automatically generating medical reports includes, though is not strictly confined to, Diagnostic Captioning (DC). In the literature, the generation of diagnostic reports is often described as a tedious task, which promotes its automatization (Yin et al., 2019a). Monshi et al. (2020) highlights several differences that make DC tasks more challenging than natural or generic image captioning methods:

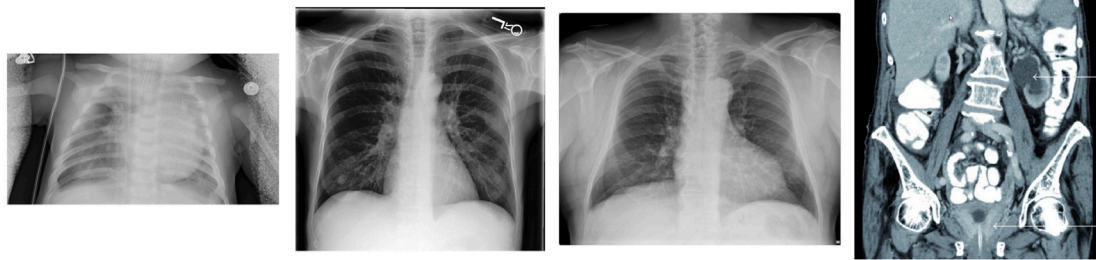


Fig. 8. Samples of (a) The Indiana University Chest X-ray Collection (IU X-ray) (Demner-Fushman et al., 2016), (b) Pathology Detection in Chest Radiographs (PadChest dataset) (Bustos et al., 2020), (c) Medical Information Mart for Intensive Care-Chest X-ray (MIMIC-CXR) (Johnson et al., 2019b), (d) Radiology Objects in Context (ROCO) (Pelka et al., 2018).

1. Medical data is often unavailable or difficult to access due to stricter regulations.
2. Diagnostic captions are usually composed of full paragraphs rather than sentence descriptions.
3. Prominent objects shown on medical images, such as healthy organs, are often not included in the report because they are not relevant for diagnosis. In contrast, DC methods focus on reporting clinically relevant information for diagnostic purposes.
4. DC methods need to be more cautious and precise: a missed object in a natural image scene does not have the same impact as a missed observable disease in a medical image.

In addition to these limitations, automatically generated reports have an inherent advantage. Autonomous tools are not meant to replace experts (in this case, physicians), but to assist them in their day-to-day tasks. In this case, the DC tools aim to provide an initial draft of a report, which can then serve as a basis for the physician to work. Subsequently, this reduces their cognitive load and prevents them from writing full reports from scratch.

4.1. Datasets for radiological report generation

Medical imaging exhibits a wide variety of image types, mainly due to the variety of acquisition technologies available, such as CT, X-rays, MRI, or PET. Fig. 8 shows some examples of these images, extracted from different radiology datasets. However, as highlighted in Beddiar et al. (2022b), X-ray-based imaging appears to be the most prominent modality of image captioning used in the literature, probably due to its low cost or the wide availability of public datasets. Hence, this review only focuses on X-ray-based radiological image datasets that can be used for report generation, more specifically, those datasets where X-ray images are paired with their corresponding report. Table 4 provides a detailed overview on the available radiology datasets.

1. **The Indiana University Chest X-ray Collection (IU X-ray)** by Demner-Fushman et al. (2016). In total, 3996 radiology reports and 8121 associated images were collected from the Indiana Network for Patient Care from the hospital picture archives.
2. **Pathology Detection in Chest Radiographs (PadChest dataset)** by Bustos et al. (2020). This is the first dataset to include Spanish-written reports, although most of them are specifically in the Valencian language. In total, it is composed of 160,868 chest radiographs from 6 different views associated with 109,931 Spanish reports, with 27% of the annotations performed manually by expert physicians and the remaining using a supervised method based on RNNs and attention mechanisms.
3. **Medical Information Mart for Intensive Care-Chest X-ray (MIMIC-CXR)** by Johnson et al. (2019a) is one of the most recent datasets, containing 371,920 chest X-rays linked to 227,943 free text reports written in English. Many variants of MIMIC-CXR have been released over the years, as outlined in Table 5, the most notable being:

- 3.1 **MIMIC-CXR-JPG** by Johnson et al. (2019b). This version contains the same CXR images as the original dataset, but in JPG format, with structured labels automatically extracted from their corresponding free text reports.
- 3.2 **MIMIC-CXR with Prior References Omitted (CXR-PRO)** by Ramesh et al. (2019). This dataset is a variation of the original MIMIC-CXR that omits references to previous radiological reports. The authors state that it addresses the problem of *a priori* hallucinated references produced by radiological report generation models.
- 3.3 **Chest Imagenome** by Wu et al. (2021). Inspired by the Visual Genome initiative (Krishna et al., 2017), the authors present a dataset containing scene graphs and nodes describing 242,072 images from the original dataset. It contains 1256 annotation combinations of relationships between 29 CXR anatomical locations, or objects with their corresponding bounding boxes, as well as their attributes structured as a scene graph. Additionally, it contains over 670,000 localized comparison relationships (for improvement, worsening, and no-change) between anatomical locations in sequential examinations. As a gold standard, it contains a manually annotated reference scene graph dataset of 500 unique patients.
- 3.4 **RadGraph** by Jain et al. (2021). This dataset contains graph versions of free text radiology reports. The development dataset consists of 500 board-certified radiology annotations from the MIMIC-CXR dataset, with which the authors trained a DL model to extract the entities and relations of the rest of the MIMIC-CXR dataset. Unlike Chest Imagenome, this dataset does not contain visual information.
- 3.5 **Making the Most of Text Semantics to Improve Biomedical Vision-Language Processing (MS-CXR)** by Boecking et al. (2022). This dataset provides 1162 pairs of image phrasing pairs, along with the location of the bounding boxes, collected in 8 different cardiopulmonary radiological findings. This dataset complements the MIMIC-CXR-JPG dataset, and comprises 1026 revised and edited bounded boxes, as well as phrases and 136 manual bounding box labels from scratch.

4. **Radiology Objects in Context (ROCO)** by Pelka et al. (2018). ROCO is a radiological image dataset created for multimodal image captioning, containing more than 81,000 images retrieved from the PubMedCentral database, automatically filtered and classified. The modalities presented in the dataset are X-rays, CT, MRI, and US, and all are accompanied by captions describing the content of the image.
5. **ImageCLEFmedical** by Rückert et al. (2023). ImageCLEF is an annual research challenge and workshop for medical image retrieval, classification, and captioning. It is part of the Cross Language Evaluation Forum (CLEF) initiative, which aims to promote the development and evaluation of information retrieval

Table 4

Available Chest X-ray radiology datasets with paired radiology reports.

Dataset name	Radiological images			Radiology reports		Source of data	Additional information
	Size	Views	Modalities	Size	Language		
The Indiana University Chest X-ray Collection IU X-ray ((Demner-Fushman et al., 2016)	8121	2	X-ray	3996	English	Indiana Network for Patient Care, United States	None
Pathology Detection in Chest Radiographs (PadChest) (Bustos et al., 2020)	160 868	6	X-ray	109 931	Spanish	San Juan Hospital, Spain	None
Medical Information Mart for Intensive Care-Chest X-ray (MIMIC-CXR) (Johnson et al., 2019a)	371 920	2	X-ray	227 943	English	Beth Israel Deaconess Medical Center Emergency Department, United States	None
Radiology Objects in COntext (ROCO) (Pelka et al., 2018)	81 000	N/A	X-ray, CT, MRI, and US	81 000	English	Retrieved from PubMedCentral	None
ImageCLEFmedical Caption Challenge (Rückert et al., 2023)	Depends on the challenge	N/A	Multi-modal	Depends on the challenge	English	Depends on the challenge	The ROCO data set is also included as part of the challenge dataset

Table 5

Variants of the MIMIC-CXR dataset.

Dataset name	Acronym	Reference	Images	Radiology reports	Dataset details
Medical Information Mart for Intensive Care-Chest X-ray 2.0	MIMIC-CXR-JPG	Johnson et al. (2019b)	377 110	227 835	Structured labels extracted from free text reports
MIMIC-CXR with Prior References Omitted	CXR-PRO	Ramesh et al. (2019)	371 920	227 943	Omits references to previous radiological reports
Chest ImaGenome	Chest ImaGenome	Wu et al. (2021)	242 072	None	1256 annotation combinations of relationships between 29 CXR anatomical locations or objects. Bounding boxes and attributes of anatomical locations or objects. Gold standard of manually annotated reference scene graph dataset of 500 unique patients.
Extracting Clinical Entities and Relations from Radiology Reports	RadGraph	Jain et al. (2021)	None	None	Graph versions of free-text radiology reports.
Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing	MS-CXR	Boecking et al. (2022)	None	None	1162 image–sentence pairs of bounding boxes and corresponding phrases, collected across eight different cardiopulmonary radiological findings

systems. The challenge is open to researchers and practitioners from academia and industry and participants can compete in several different challenges, such as image annotation, image retrieval, and classification. The results are evaluated using standardized parameters. The ROCO data set is also included as part of the challenge dataset.

Although the Indiana University Chest (ChestIU) X-ray dataset is one of the first and most used in the literature (Beddiar et al., 2022b), MIMIC-CXR is the largest dataset. ImageCLEFMedical or ROCO are also very useful datasets, as they contain multimodal and heterogeneous data, which more accurately resembles real-world data.

4.2. Evaluation methods

Automated captioning methods have traditionally been evaluated using machine translation and text summarization approaches (Pavlopoulos et al., 2022). These metrics are based on the comparison between the proposed caption and a ground-truth label. In the context of radiological reports, the ground-truth is the report redacted by a certified radiologist. Most of the NIC and MIC models presented in this survey are evaluated using this approach. From a wider perspective, MIC evaluation methods can be grouped on the basis of the level of subjectivity as quantitative and qualitative, the first being the most common. Quantitative methods can be further classified, based on the evaluation goal, as Natural Language Generation (NLG) and Clinical Efficacy (CE).

4.2.1. Natural language generation evaluation methods

NLG evaluation methods are based on word overlap metrics, which aim to estimate the similarity between the generated and ground-truth captions. Similarity in this context is based on how many words or *n-graphs* (phrases of *n* consecutive words) are shared between the generated and real captions. NLG metrics include:

1. **Bilingual Evaluation Understudy (BLEU)** by Papineni et al. (2002). BLEU measures the position-independent matches of *n-grams* and introduces a brevity penalty (BP) to penalize short generated captions. The closer the BLEU value is to 1, the better the model performs. Usually, BLEU-n is the preferred nomenclature, as it also specifies the *n* words considered in the *n-grams* method.
2. **Metric for Evaluation of Translation with Explicit Ordering (METEOR)** by Banerjee and Lavie (2005). METEOR was developed to evaluate the correlation between human- and machine-generated captions, at the sentence level. It extends BLEU-1 employing the harmonic mean between Precision and Recall, F_β score biased toward recall. The F_β score is the generalization of the F1 score, where $\beta = 1$, which is then reduced by up to 50% if there are no overlaps *n-grams* between the machine and the human reference descriptions. For machine translation, values higher than 0.6 are considered better than human performance, since even two humans are not likely to achieve a perfect translation match. For this reason, a BLEU score closer to 1 is

often perceived as unrealistic and could be an indicator that the trained model is overfitted.

3. **Recall Oriented Understudy for Gisting Evaluation (ROUGE)** by Lin (2004). This metric was introduced to estimate the quality of a summary, but also evaluates the *n-gram* overlapping between two descriptions. Many variants have been published, based on different *n-gram* lengths, stemming, or stopword removal (Graham, 2015), but baseline ROUGE-L is still the most widely used metric in MIC (Pavlopoulos et al., 2022).
4. **Consensus-based Image Description Evaluation (CIDER)** by Vedantam et al. (2015). Designed to evaluate image description, it measures the cosine similarity between *n-gram* TF-IDF representations of both ground-truth and generated captions. Beddiar et al. (2022b) notes that this metric considers grammar, saliency, and accuracy.
5. **BERTScore** by Zhang et al. (2019). BERTScore is an automatic evaluation metric for text generation. This metric computes the similarity score for each token in the reference sentence using contextual embeddings. The authors show that it correlates better with human judgements and provides a stronger model selection performance than other existing metrics. In fact, the aforementioned ImageCLEFmedical challenges from recent years include BERTScore as part of their evaluation metrics (Ionescu et al., 2023).

4.2.2. Clinical efficacy evaluation methods

Measured based on word overlapping, such as those reported in Section 4.2.1, do not necessarily capture clinical correctness. Clinical Efficacy (CE) methods have been widely used, in conjunction with NLG methods, to overcome this problem. Diagnostic information is obtained by extracting a set of medical terms from the generated caption and comparing them with the ground-truth caption. These reference terms may be generated by humans (such as those in the IU X-ray data set), or by a system labeler (such as in the Medical Text Indexer (Mork et al., 2013) or the CheXpert labeler (Irvin et al., 2019)). These evaluation metrics englobe conventional classification metrics such as accuracy, precision, recall, F1-score, or AUC-ROC.

4.3. Medical report generation methods

NIC methods presented in Section 3 are often applicable to diagnostic captioning from medical images. However, as previously noted in this survey and as remarked by Pavlopoulos et al. (2022), there are two notable differences. First, unlike NIC methods that simply describe the content of an image, DC methods focus on highlighting diagnostically relevant aspects of the image. Second, the reports generated are often quite similar between studies of different patients, where retrieval-based methods have shown performance comparable to those based on deep ED architectures.

Several recent surveys and state-of-the-art reviews have been published focusing on the topic of automatic radiology report generation (Pavlopoulos et al., 2019; Monshi et al., 2020; Ayesha et al., 2021; Pavlopoulos et al., 2022; Beddiar et al., 2022b; Messina et al., 2022). Pavlopoulos et al. (2019) presented one of the first surveys on this subject. This survey presented an initial categorization of the existing approaches, while also noting the difficulty of comparing the methods among themselves, since each of them made use of different datasets. This issue could be caused by the lack of well-established datasets at the time, opposite to NIC. Subsequent publications, such as Monshi et al. (2020) and Ayesha et al. (2021), focused on DL techniques to generate radiological reports from scratch. These works do not consider retrieval-based techniques, while Pavlopoulos et al. (2022) extended their initial publication and provided a survey, highlighting the shortcomings of existing DC datasets and evaluation measures. Approximately at the same time, Beddiar et al. (2022b) conducted a more thorough and technical analysis of the state-of-the-art. Messina

et al. (2022) also published their survey, which extensively focused on explainable AI approaches for DC, which is also slightly addressed in the work by Nazir et al. (2023). Finally, Shamshad et al. (2023) explored the use of transformers in the context of medical imaging.

According to Beddiar et al. (2022b), DC methods can be grouped according to how the final diagnostic descriptions are generated. Simple image descriptors (such as those generated with early approaches) or higher-dimensional features (generated with more recent and deeper approaches) are often extracted from medical images and then coupled with one or more text generation techniques. The taxonomy proposed by Beddiar et al. (2022b) divides the DC methods into four categories: retrieval-based approaches, template-based approaches, generative models, and hybrid models. This categorization serves as the foundation for organizing and presenting the work reviewed in this paper. Due to the scope of the paper, the main focus is on the application of DC techniques for CXR images, but encompasses different types of DC approach, as medical image-related methods follow similar diagnostic and grammatical objectives.

4.3.1. Retrieval-based methods

As in the case of natural images, medical images also tend to use the same caption for different images. Retrieval-based methods query a set of visually similar images from a database and assign to the novel image the top, or a combination, of the top *k*-captions from the retrieved set following a set of predefined rules and schemes (Ayesha et al., 2021). Even in their simplest form, where reports of the most similar images are reused as is, retrieval-based models have been reported to have higher recall than other approaches (Liu et al., 2019). Depending on the clinical context, such as in screenings, recall is often a preferred metric as it aims to detect most of the true positives. However, precision is also critical in different scenarios, such as the validation of clinical impressions.

Wang et al. (2018b) proposed a transfer learning for a multi-label classification scheme in combination with a retrieval-based method to generate the corresponding caption. Similar images are retrieved according to the color and texture features. Then, the captions of the top 3 matches are selected and combined to generate a new caption. This method achieved a final mean BLEU score of 0.25. Charalampakos et al. (2021) presented a retrieval approach based on *k*-Nearest Neighbors (*k*-NN) in addition to their previous work (Karatzas et al., 2020), where images are recovered from the training set based on the cosine similarity between the image embeddings.

Other methods, such as Syeda-Mahmood et al. (2020) proposed a domain-aware approach, exploiting a feature pyramid for DC retrieval. This method first learns fine-grained descriptions of image findings and then uses the extracted features to retrieve and customize similar reports from a large database. Although retrieval-based methods have shown very promising results for DC, they are still limited in the preparation of large databases, since they are incapable of representing the heterogeneity of pathologies present in real-world data (Chen et al., 2020).

4.3.2. Template-based methods

These methods follow the same concept as the methods presented in Section 3.2.2, where the generation of caption templates is performed following some specific rules. In the context of DC, these templates are filled with the text that describes the diagnostic findings. Previous work, such as Kisilev et al. (2015), applied DC methods for the generation of breast radiology reports. The image descriptors were first extracted from the image using a Support Vector Machine model. The extracted semantics descriptors are then projected in the predicted text case report.

Language templates are often coupled with deep encoder architectures to extract higher-order features from the image and then generate the corresponding report based on templates. In fact, the automatic

report generation task is often divided into two tasks: image classification and text generation. Harzig et al. (2019) used a deep CNN to analyze gastrointestinal examinations and detect diseases, while the generated description was obtained by filling in a textual template with the corresponding findings. Pino et al. (2021) proposes an even simpler strategy, by implementing a multi-label classifier of the 13 abnormalities defined by the CheXpert labeler (Irvin et al., 2019). The authors manually define a positive and negative description given by the presence or absence of each label in the image, which are then all concatenated to generate the final report. This approach was shown to be more clinically accurate than other approaches based on DL or attention. However, since it only describes the presence or absence of a certain disease, the generated description lacks information related to the orientation, relative location, or extent of the disease.

Template-based methods are simple and follow the correct grammar, but rely on preset visual concepts that may limit the flexibility and variety of the results, as noted in Ayesha et al. (2021). Furthermore, Beddiar et al. (2022b) noted the difficulty in filling out templates for abnormal findings, in contrast to normal screenings whose reports tend to be very similar between different patient studies. They argue that abnormalities should be explained and localized in the image and that the generated templates should be adapted to include descriptions of such abnormalities.

4.3.3. Generative methods

In this context, generative methods are those in which a deep neural network is trained from end to end to learn image-to-caption mapping. Due to the rise of DL methods and the constant progress in the area, most of the methods studied with regard to the generation of medical reports fall into this category. Thus, this classification is then decomposed into different architectures and learning strategies, which are explored in the following subsections.

Encoder–decoder architectures. The idea of ED architectures for the generation of medical reports is the same behind the ED architectures presented in Section 3.3.1: a CNN encodes the input image, while a RNN decodes the generated caption. Several works even adapted the *Show and Tell* model for DC, showing promising results (Pelka et al., 2017; Tsuneda et al., 2021). Other works like Sun et al. (2019) have focused on improving encoder–decoder training on smaller datasets, where they adopted a Fully Convolutional Network (FCN) and multi-label classification layers as encoders and an LSTM as decoder.

From a general perspective, most authors that exclusively implement ED architectures often introduce well-established CNN architectures for the encoder and RNNs such as LSTM and GRUs as the decoders, along with certain variations to more appropriately fit the DC task (Monshi et al., 2020; Ayesha et al., 2021; Pavlopoulos et al., 2022; Beddiar et al., 2022b). Transfer learning is also commonly used in these architectures, where the encoder is often trained on larger datasets of natural images. As noted in Beddiar et al. (2022b), these methods seem to perform well, but they still have several limitations. Most importantly, current sentence generation evaluation metrics are not yet capable of capturing the change in meaning in a sentence when a punctuation mark or a negation appears (Singh et al., 2019; Zeng et al., 2020b; Xue and Huang, 2019). In addition, desirable reports should describe not only the normal and abnormal findings, but also the location and characteristics of such abnormalities (Li et al., 2018; Yin et al., 2019a; Ouyang et al., 2020), which the ED approaches cannot describe (Ambati and Dudyala, 2018; Xie et al., 2019; Beddiar et al., 2022b).

Attention-based models. Attention-based models for NIC were reported in Section 3.3.3. DC also benefited from the surge in popularity of attention-based methods. Zhang et al. (2017b) is one of the first works to introduce this concept into the medical domain to generate diagnostic captions for images of bladder cancer. They present MDNet, a CNN encoder whose features are attended by an auxiliary attention

sharpening module and later fed to an LSTM. Other works, inspired by the NIC method *Show, Attend, and Tell* (presented in Section 3.3.3), introduce visual attention in their CNN-RNN architecture (Xu et al., 2019). However, these models barely improve the results achieved by generic NIC models adapted for diagnostic captioning, such as Karpathy and Fei-Fei (2015), or Xu et al. (2015).

More recently, transformers have also been included as part of both the encoder and decoder (Pavlopoulos et al., 2022; Beddiar et al., 2022b; Shamshad et al., 2023). One of the first approaches was to use CNNs to encode full-size images in smaller encoding patches (Chen et al., 2020). These patches are then fed to a transformer-based architecture to generate the final medical report. Other approaches opt for Visual Transformer architectures (ViT) (Dosovitskiy et al., 2020) as the basis for the generation of radiology reports (Nicolson et al., 2021; Zhou et al., 2022; Mohsan et al., 2022; Wang et al., 2023). Nicolson et al. (2021) combined a ViT encoder with PubMed-BERT (Gu et al., 2021), a pre-training transformer model with PubMed articles, and domain-specific self-supervised training strategies. The combination of both models gave them a third place in the ImageCLEFMed Caption 2021 challenge, achieving a score only 0.078 worse than the winning proposal.

With the growing interest in explainability in deep learning models, new approaches are increasingly dedicated to addressing this aspect, extending their focus to other models intended to tackle and incorporate it within their proposals (Messina et al., 2022; Nazir et al., 2023). Hou et al. (2021) introduced attention to identify regions of the input image and demonstrate where the model is focusing to generate the final report.

Reinforcement learning models. Instead of using regular supervised learning approaches, some proposals rely on Reinforcement Learning (RL) as their optimization strategy. One of the main reasons to use RL is its flexibility to optimize nondifferentiable reward functions (Messina et al., 2022). This lack of constraints with respect to differentiability allows researchers to be more creative in defining and implementing new reward functions. For example, Liu et al. (2019) designates two reward functions, aligned with the objectives of DC: a natural language reward function based on the value achieved by the model in the CIDEr (Section 4.2), and a Clinically Accurate Reward that measures the diagnostic quality of a report with respect to a reference to ground-truth using CheXpert. Other works simply optimize the model value for the CIDEr and BLEU-4 values during training (Xiong et al., 2019; Jing et al., 2020).

Xiong et al. (2019) was one of the first approaches to merge attention and reinforcement learning, proposing a reinforced transformer for DC. The authors coupled a pre-trained DenseNet model to identify ROIs in images and a transformer to extract visual features and generate diagnostic captions. Later, Miura et al. (2020) also used DenseNet as an encoder but instead used a Meshed-Memory Transformer (Cornia et al., 2020), originally intended for NIC, as a decoder. Miura et al. (2020) argue that the architecture proposed by Cornia et al. (2020) is more effective for DC than traditional approaches based on RNN and transformers. Miura et al. (2020) also proposed two reward functions using self-critical RL as an optimization strategy, promoting coverage and consistency in reports generated with respect to their ground-truth reports. The results achieved by these reward functions are then combined with the semantic equivalence metric provided by the BERTScore. The authors found that this approach achieved better performance in the generation of radiology reports in terms of clinical metrics.

Compositional models. Compositional models consist of a set of independently trained models, each of them being responsible for a subtask. Chelaramani et al. (2020) defined the task of generating medical reports as a three-fold task: (1) prediction of the coarse-grained broad-disease category; (2) prediction of the fine-grained disease subcategory; and (3) generation of a textual diagnosis. They compared

the performance of the final model when the three submodels were trained independently or simultaneously, with the latter achieving better results. Similarly, Zeng et al. (2018) decomposed the US captioning problem into subtasks. First, a CNN-based organ classifier model is trained. Then, each organ-specific CNN-based encoder is trained together with an organ-specific LSTM to generate the final caption.

Attention-based models can also be used as building blocks for compositional architectures. Jing et al. (2018) proposed a multitask hierarchical model with co-attention for predicting keywords and generating long paragraphs. They use a CNN for tag prediction, where tags are defined as word embeddings which are then fed together with visual features to the co-attention module. This resulted in two semantic and visual context vectors that are concatenated and passed to a hierarchical LSTM: a sentence LSTM and a word LSTM. Variations of their work have also been published with other encoding architectures (Wang et al., 2018a; Liu et al., 2019). Similarly, Yuan et al. (2019) proposed a multiview approach, using CNN ResNet for classification and a hierarchical attention-based LSTM for caption generation. It reported a state-of-the-art result, surpassing those achieved by Jing et al. (2018) and Li et al. (2019a). In these cases, each submodel was trained separately.

Dense captioning models. As stated in Section 3.3.4, dense captioning models aim to address the problem of locating and provide fine-grained descriptions of ROIs in an image. In the context of DC, accurate detection and description of ROIs in the image are essential to provide a complete and detailed report. Contrary to what Pavlopoulos et al. (2022) suggest, dense captioning can be useful for medical practitioners, residents, and other health professionals, and even people with not as deep knowledge of medicine as physicians. It can also serve as a formative interactive tool to provide more descriptions along with associated visual regions, rather than a long and wordy diagnostic caption of the whole input. Furthermore, as highlighted in Hossain et al. (2019), dense captioning systems tend to be more objective than full image captioning approaches, which is aligned with the goal of diagnostic captioning and the generation of radiology reports.

Diagnostic reports are inherently long and describe several key biomedical features, remarking the need for models to appropriately capture the strong semantic knowledge for medical image-text report mapping. In fact, dense captioning, which is obtained through ROI algorithms like those used in object detection, can be combined and processed to generate complete radiology reports. For example, Li et al. (2019c) were among the first works to address report generation from dense captions. The proposed architecture, *Vispi*, is composed of DenseNet-121 (Huang et al., 2017) for binary classification and Grad-CAM (Selvaraju et al., 2017) for unsupervised localization. If a disease is localized, the region feeds a CNN coupled with an attentive LSTM to generate a report. Otherwise, for images labeled as *normal*, a simple attentive LSTM produces the report without including any visual clues.

Later, Zeng et al. (2020a) proposed a two-phased caption generation model for ultrasound images based on object detection. First, a faster R-CNN is trained to detect and classify diseases on ultrasound images. Then, the feature vectors of the detected ROIs serve as input for an LSTM caption generation model, generating a description for each ROI. This work was later extended by Zeng et al. (2020b), where they proposed a Semantic Fusion Network (SFNet) as an ROI detection proposal. Then, the extracted feature vectors is passed to an LSTM, which generates the full description for the ultrasound images. Although the final task is to generate full diagnostic captions, the intermediate step is to obtain these dense captions of the recognized objects with the coordinates of the detection of the bounding box in the image. This information can then be fed into the report generation model so that the descriptions contain notions of location and orientation of injuries.

4.3.4. Hybrid methods

Template-based and generative models. Since medical reports usually follow a static structure, template-based methods initially seem to be a good approach to address DC. However, as stated in Section 4.3.2, these approaches may not be flexible enough to fit the variety of results. Therefore, several works have aimed to merge template-based and generative-based approaches to combine the benefits of both methods. For example, Gale et al. (2019) argued that automatic medical report generation of frontal pelvic X-rays, limited to the task of classifying hip fractures, is such a narrow task that it can be simplified to use only two sentence templates. They encoded the input image with DenseNet and classified it as positive or negative. Positive instances are then fed into an LSTM with an attention mechanism over the image encoding to fill the template slots. Other works further extend and try to improve this hybrid method (Han et al., 2018; Xie et al., 2019; Han et al., 2021). In particular, Han et al. (2021) propose to employ an adversarial graph network for semantic segmentation to detect abnormalities, and a unified template to report findings.

Retrieval-based and generative models. Section 4.3.1 outlined the advantages and limitations of retrieval-based models. Similarly to hybrid template-based methods, retrieval-based approaches can be coupled with generative models to better adapt previously generated similar reports to new input data. This combination suppresses the limitations exhibited by template-based methods, as these hybrid methods can introduce the notion of generalization over unseen data (Beddiar et al., 2022b). Some methods even introduce RL as an optimization strategy to decide when to stop recycling previous reports or generate new ones (Li et al., 2018; Xiong et al., 2019).

Li et al. (2019a) proposed a novel Knowledge-Driven Encode, Retrieve, Paraphrase (KERP) model. This proposal splits the task into three phases: encode, retrieval, and paraphrase. First, a modified version of a Graph Transformer (GRT) encodes an input image to generate a knowledge graph of abnormality attributes. Then, this graph is used to guide the retrieval process to capture the most pertinent templates. Finally, the template sequence is further refined through the paraphrase module. This last module is composed of another variant of GRT for graph and sequence input, and sequence output, where the template words and the encoded knowledge graph are the attended elements. This work is compared with other attention-based model baselines, with the IU X-ray dataset and a private Chinese dataset outperforming all other models in the BLEU-1,2,3,4 scores.

5. Discussion

Image captioning has gained popularity in recent years, especially following the surge of DL and attention-based models, by bridging the gap between two major research fields: computer vision and natural language processing. As highlighted in Section 3, automatic image caption generation is a particularly challenging task, especially in the medical domain, where expert knowledge plays a key role. This Section presents an initial analysis based on the revised literature on image captioning methods, focusing on Natural Image Captioning and Medical Image Captioning.

5.1. Image captioning and medical report generation

NIC has the particularity that the developed models need to understand the complexity and relationships among numerous classes. Early work was characterized by two stances: retrieval and template-based methods. Both approaches require a time-consuming process of hand-crafting features. Retrieval-based methods required methods to evaluate the similarity between images, while template-based methods demanded ontologically fixed templates with predefined slots for caption generation. Although these methods generate syntactically correct captions, they cannot generate image-specific captions with semantic correctness.

The progress in DL has revolutionized image captioning methods based on templates and retrieval, surpassing older hard-engineered approaches. Deep CNNs and increased computational power have played a crucial role in this advancement. Although initial methods relied on shallow models and manual feature engineering, DL approaches excel at image captioning. Fig. 6 depicts the taxonomy of DL-based image captioning. Of the methods reported in the taxonomy, supervised learning of curated captions is the most adopted learning strategy in the field, although unsupervised methods based on GANs and RL have also gained interest. Although unsupervised approaches can generate diverse image captions, supervised learning models still achieve superior results. Notable architectures include the Encoder–Decoder model, inspired by machine translation, which learns image features in an end-to-end manner using a CNN as the encoder and a LM as the decoder.

Compositional architectures and attention-based methods have also made significant contributions to NIC. The first employ multiple independently trained building blocks, such as a CNN for image feature extraction, an LM for generating candidate captions, and a deep ranking-based model to select the best result. Moreover, attention-based methods have gained popularity by considering spatial aspects of input images and dynamically focusing on relevant regions during caption generation. These methods often incorporate attention mechanisms into the ED scheme, such as the encoder, decoder, or both. Notable attention-based models include *Show, Attend, and Tell* (Xu et al., 2015), which integrates attention mechanisms on visual features extracted by a CNN and then passed to an LSTM-based decoder. Another relevant work in this area is that by Anderson et al. (2018), for Bottom-Up and Up–Down Attention, incorporating hard and soft attention mechanisms through the use of faster R-CNN and LSTM architectures. Researchers have also explored the integration of GCNs and MHA mechanisms to enhance attention-based models, leading to improved semantic information and spatial relationships in generated captions. The continuous exploration of novel techniques and architectures further drives advances in image captioning.

Dense captioning is another variant of NIC in which a description is generated for each ROI identified in the image. These methods improve the interpretability and utility of visual data by enabling fine-grained descriptions and facilitating richer visual understanding in a variety of applications and domains. However, these approaches present a series of challenges, such as accurately recognizing target regions and handling overlapping ROIs, as highlighted by Shao et al. (2022).

5.2. Medical image captioning

Medical Image Captioning comprises three main aspects: availability of existing datasets, evaluation methods, and methods for generating medical reports. By addressing these aspects and studying the retrieved resources, we gain insight into current advancements, limitations, and future directions of medical image captioning, as well as their applications in the context of radiology.

5.2.1. Available datasets

Fig. 9 depicts the evolution in time of the citations of the MIC datasets reported in Section 4.1. The ChestIU X-ray Dataset gained early popularity and widespread use upon its publication, as it was among the first datasets specifically focused on CXR images with attached reports. Its usage reached a peak in 2021, but is now slowly decaying and being replaced by newer datasets: ROCO, PadChest, and MIMIC-CXR. The increase in citations for ROCO, a dataset characterized by multimodal medical images, highlights the growing interest in developing multimodal models capable of integrating different modalities, similar to the knowledge and expertise of practitioners. PadChest and MIMIC-CXR are fairly similar datasets, with around 160,868 and 371,920 CXR images, respectively. PadChest is the only known dataset of its size that includes Spanish-written reports, although most of them are written

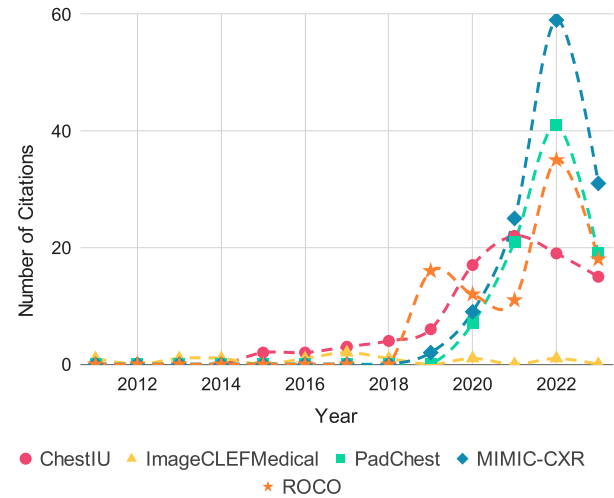


Fig. 9. Evolution in time of citations on the main MIC datasets. ChestIU, MIMIC-CXR, and PadChest citations were based on the number of citations in PubMed of the main reference publication. ImageCLEFMedical count was based on a PubMed search of ImageCLEFMedical since it is a yearly proposed challenge. ROCO citation count was based on Google Scholar search as the main referencing publication was not indexed in PubMed. Note: 2023 only includes publications between January and June 2023.

in the Valencian language. However, it provides opportunities for language-specific research and analysis of MIC methods. MIMIC-CXR, on the contrary, consists of English-written reports, which contributes to its wider accessibility and usability.

Several variants of the MIMIC-CXR dataset have been released, including MIMIC-CXR-JPG, which provides the same images in JPG format with structured labels extracted from the reports. Chest Imagenome and RadGraph are very promising data sets that contain entities and relations in the form of graphs extracted from radiology reports. The chest Imagenome contains visual information of the anatomical regions present in the images. Moreover, MS-CXR offers bounding box image–phrase pairs for cardiopulmonary findings, complementing the MIMIC-CXR-JPG dataset for the dense captioning task. Lastly, although no citations on CXR-PRO were found, it may be useful to address the hallucination problem that is fairly common in MIC. All of these datasets provide diverse resources for training and evaluating MIC models, covering different image formats, report structures, and graph-based annotations.

5.2.2. Evaluation methods

Evaluation methods for MIC can be categorized into quantitative and qualitative, as described in Section 4.2. Quantitative methods include Natural Language Generation (NLG) methods and Clinical Efficacy (CE) methods. NLG methods assess the similarity between the generated report and ground-truth captions using word overlapping metrics such as BLEU, METEOR, ROUGE, CIDEr, and BERTScore. These metrics capture the linguistic quality and similarity of the captions. BERTScore, in particular, is gaining popularity as an evaluation metric for MIC due to its ability to overcome the limitations of previous methods such as BLEU and METEOR. Current evaluation methods struggle to capture changes in meaning caused by punctuation or negation. Using contextual embeddings and evaluating similarity at the token level, BERTScore captures semantic meaning and context more effectively. However, CE methods focus on evaluating the clinical correctness of captions by comparing medical terms extracted from the generated and the reference captions, based on widely established models like the Medical Text Indexer or the CheXpert Labeler. CE methods often use conventional classification methods such as accuracy, precision, recall, F1 score, and AUC-ROC. Combining NLG and CE methods leverage a comprehensive evaluation framework that can be applied to the models

Table 6
Summary of the revised MIC approaches (1).

Method	Architecture	Description	Datasets
TieNet (Wang et al., 2018a)	Retrieval-based	<ul style="list-style-type: none"> Enables multi-label classification to detect diseases from images A CNN-RNN architecture is used to generate reports from images 	IU-XRay
Charalampakos et al. (2021)	Retrieval-based	<ul style="list-style-type: none"> Relies on a K-NN to detect the existing captioned images most similar to the input Captions are generated from the combination of the captions of the most similar images 	ImageCLEF 2021
Syeda-Mahmood et al. (2020)	Retrieval-based	<ul style="list-style-type: none"> Introduces Fine Finding Labels to generate the description of images Bases the generation of new reports on combining existing reports labeled with the FFLs detected in the image 	IU-XRay
Pino et al. (2021)	Template-based	<ul style="list-style-type: none"> It comprises two parts, an image classification module and a template report generation module The CNN first performs multilabel classification to detect abnormalities in the image Two templates are used: one indicating the presence and one indicating the absence of a given abnormality The final report is generated by concatenating all the absence/presence sentences 	IU-XRay MIMIC-CXR
METransformer (Wang et al., 2023)	Generative method	<ul style="list-style-type: none"> Uses “expert tokens” to capture fine-grained visual patterns Expert tokens are included both in the encoder and the decoder Addresses the difficulty of focusing on correct regions of the image 	IU-XRay MIMIC-CXR
MDNet (Zhang et al., 2017b)	Generative method	<ul style="list-style-type: none"> Comprises an image model and a language model The image model feeds a set of convolutional features and an image embedding to the language model The LSTM, which includes an attention mechanism, maps sentence words to image pixels to generate a final report 	BCDIR
Chen et al. (2020)	Generative method	<ul style="list-style-type: none"> The proposed model comprises three components: a visual extractor, an encoder and a decoder Relational memory is included in the decoder to learn from the previous patterns and achieve better performance overtime A sequence-to-sequence approach is followed to generate the reports 	IU-XRay MIMIC-CXR
RATCHET (Hou et al., 2021)	Generative method	<ul style="list-style-type: none"> Is based on a CNN-RNN transformer An attention mechanism is used to accurately detect regions of interest in the image For each part of the generated report, RATCHET indicates on which region of interest is focusing 	MIMIC-CXR
Liu et al. (2019)	Generative method	<ul style="list-style-type: none"> Employs a hierarchical strategy, using a CNN for image encoding, and two subsequent RNN for sentence and word decoder, respectively The proposed model is finetuned to consider quality aspects such as readability and clinical accuracy. 	IU-XRay MIMIC-CXR
RTMIC (Xiong et al., 2019)	Generative method	<ul style="list-style-type: none"> Relies on the accurate detection of regions of interest The relations between the different regions of interest are also encoded, providing semantic information An attention mechanism is used to generate the final report from the regions of interest and the extracted semantic information 	IU-XRay
M ² Trans (Miura et al., 2020)	Generative method	<ul style="list-style-type: none"> Introduce two novel rewards to generate better quality reports, focusing on two aspects: completeness and consistency. Reinforcement learning is employed to optimize the rewards It employs M2Trans as the base image captioning model 	IU-XRay MIMIC-CXR
Jing et al. (2020)	Generative method	<ul style="list-style-type: none"> Presents a multi-task learning framework that jointly predicts abnormalities in the image and generates paragraphs for their description Use co-attention to accurately locate regions of interest and the relations between them A hierarchical LSTM is used to generate long descriptive paragraphs instead of short sentences 	IU-XRay

studied. Most published works report multiple metrics, as it is widely acknowledged that no single metric is adequate or sufficient to assess the quality and clinical relevance of diagnostic captioning.

5.2.3. Diagnostic captioning methods

The field of automatic interpretation of medical images has made significant progress. Several surveys and reviews have been conducted to explore automatic diagnostic captioning of images, provide information on different methodologies, and highlight existing challenges. Tables 6 and 7 summarize and provide an overview of the characteristics of the revised methods.

Sections 4.3.1 and 4.3.2 discussed the advantages and limitations of template- and retrieval-based methods, respectively. In the case of DC, since radiology reports are often repetitive and very structured, these methods perform surprisingly well. For example, retrieval-based models, even when using unchanged reports from the most similar image, exhibit a higher sensitivity than other approaches, making them suitable for screening scenarios in which detecting true positives is crucial (Wang et al., 2018b). However, they still face challenges in preparing large databases or well-defined templates that can capture the heterogeneity of abnormalities and pathologies in real-world data.

Section 4.3 discussed the different generative approaches and architectures employed in DC, where the term *generative* refers to the ability of the model to produce fully novel, realistic and accurate reports that resemble training data. These approaches have been gaining significant interest in the field, particularly those based on DL and attention methods, with the majority of revised publications falling into this category. Generative methods were further classified according to the architectures and learning strategies used. Encoder–Decoder (ED) architectures based on CNNs and RNNs have been widely adopted for caption generation, along with transfer learning approaches to overcome data availability, generalization, and robustness problems. Several studies have identified limitations of these architectures, putting an emphasis on the fact that ED architectures lack details about the location and characteristics of abnormalities, which ideal reports should include (Ambati and Dudyala, 2018; Xie et al., 2019; Beddiar et al., 2022b).

Following ED, attention-based models, inspired by the success of the *Show, Attend, and Tell* approach in NIC, have also been applied to DC. Transformer-based architectures, both CNN-based and visual Transformer (ViT) encoders, are among the most relevant attention methods. Attention has also impacted disease localization studies for report generation, which in a certain way correspond to NIC’s dense

Table 7
Summary of the revised MIC approaches (2).

Method	Architecture	Description	Datasets
Yuan et al. (2019)	Generative method	<ul style="list-style-type: none"> Exploits the information given by multi-view images Multi-view information is synthesized using a sentence-level attention model Include medical concepts in the decoder with word-level attention to ensure that the medical-related contents are correct 	IU-XRay
SFNet (Zeng et al., 2020b)	Generative method	<ul style="list-style-type: none"> Includes semantic and grammatical information within the generation process Employ a lesion area detection model to accurately detect regions of interest within the image 	IU-XRay
KERP (Li et al., 2018)	Hybrid method	<ul style="list-style-type: none"> Comprises three modules: encode, retrieve and paraphrase In the encode module, the image is encoded via a CNN to generate an abnormality graph that indicates the potential abnormalities present in the image and the relationship between them. The retrieve phase decodes the abnormality graph in a set of templates sequences The paraphrase module refines the templates with enriched details. 	IU-XRay

captioning for medical images. These approaches, where a description is coupled with visual detections of where diseases are found or are progressing, increase the explainability and understanding of the reasoning process behind these black-box models.

Other learning strategies for DC methods reported in this survey include RL-based algorithms. They are a more flexible alternative to supervised learning methods since they allow novel, nondifferentiable reward functions to be included in the optimization process. Relevant progress has been achieved by combining attention mechanisms with RL, achieving improved performance in terms of clinical metrics. Other promising research directions aim at implementing compositional architectures, where independent models are trained for different subtasks. In addition, when these strategies are coupled with attention-based mechanisms, they are reported to show state-of-the-art performance. Often, when these architectures are studied, authors evaluate the performance of the final model under different training strategies: training each model individually or training the full architecture jointly.

Approaches that combine earlier template- and retrieval-based methods with more advanced and deeper generative methods have also gained interest in the field. Hybrid methods aim to compensate for the limitations of baseline methods by introducing a higher generalization capacity, flexibility, and patient-specific analysis characteristic of generative approaches. Hybrid approaches can also incorporate other training strategies, such as RL, to determine when to reuse existing reports or generate new ones from scratch. A novel approach called KERP, proposed by Li et al. (2019a) involves several modified Graph Transformer models to encode the image, retrieve pertinent templates guided by a knowledge graph, and refine the sequences through a paraphrasing module. This approach outperformed attention-based models in terms of BLEU score. Hybrid methods provide alternatives to improve DC performance while addressing the challenges of generating clinically accurate and informative reports.

Fig. 10 shows the performance of the models reviewed in Section 4.3 on MIMIC-CXR. One of the first significant results of the comparison is the high similarity between the values of the BLEU and ROUGE metrics of the retrieval-based methods studied. This similarity implies that there is a good trade-off between precision and recall. However, it is also worth mentioning that in the case of TieNet, the CIDEr score is significantly low. A similar phenomenon occurs in the case of Pino et al. (2021), where the CIDEr score is also below 0.25. Since both retrieval and template-based models follow a sequence-to-sequence approach to generate the final reports without including any additional semantics, this may be the reason behind these low scores, since CIDEr is the only metric that considers the quality of the content and its meaning. In the same vein, it is worth noting that RTMIC, Jing et al. (2018), and SFNet achieve better values at this metric than the rest of the generative models. This is directly related to the philosophy behind these models, since they actively include semantic information within the generation process by modeling the relations between the different regions of interest. Finally, it is worth noticing the remarkable performance achieved by METransformer, both in terms of the ROUGE

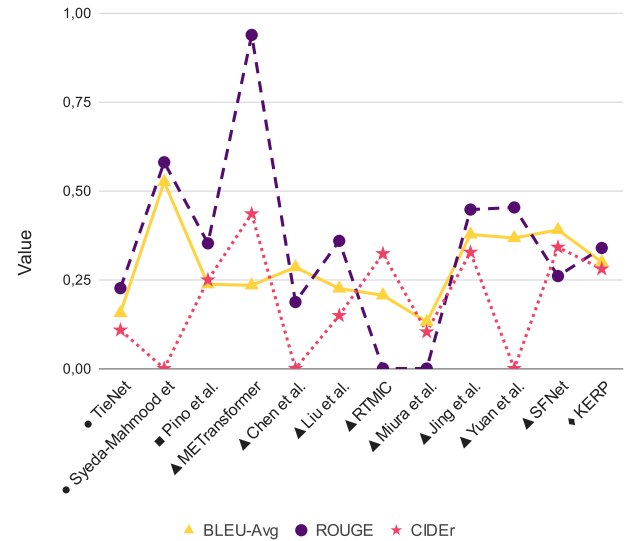


Fig. 10. Performance comparison of the MIC models reviewed in Section 4.3 on MIMIC-CXR. For each method, their average BLEU score, CIDEr and ROUGE metrics are provided. The following notation is employed after the name of each method to identify its type: ● (retrieval-based), ■ (template-based), ▲ (generative), and ◆ (hybrid). Values of 0 denote that there were no available results for the given metric.

metric and CIDEr. This model includes “expert tokens” to capture fine-grained visual patterns, which are then included in the decoder. As a result, the generated reports are almost identical to those written by a practitioner, achieving good performance not only in terms of accurately detecting abnormalities within the images (as reflected by the high ROUGE score), but also in terms of quality.

5.3. Limitations

Although Medical Image Captioning methods have shown promising results, they still exhibit certain shortcomings that hinder them from reaching their full potential. First, Beddiar et al. (2022b) argue that MIC methods should not only describe anomalies in images, but also provide explanations and precise locations for these anomalies. This is particularly relevant in the radiology domain, as accurate identification and localization of pathologies can significantly affect diagnosis and treatment planning. However, existing MIC models often lack the ability to incorporate detailed information on the location and characteristics of abnormalities, which good reports should include. Dense captioning methods aim to solve this issue, but additional efforts are required to overcome the difficulty of recognizing all target regions for each visual concept and multiple overlapping ROIs.

Second, it is important to note that DC methods exhibit a different behavior compared to NIC methods. In NIC, the goal is to describe every object present in the image, including even the most obvious

and normal objects, while DC methods aim to include only the relevant pathological aspects that can be derived from the images. This implies that certain normal objects, such as organs, should not be included in the final report. Instead, the focus should be on highlighting the improvement, worsening, localization, and urgency of the disease. This divergence of objectives between the NIC and the DC evidences the need for specialized approaches tailored to the needs of the medical domain.

Another key limitation of DC is the problem of hallucination. Hallucination refers to the phenomenon in which generated captions contain information that is not actually present in the image. In DC, models are trained with reports that contain references to previous reports, which often causes their outputs to contain hallucinated references to non-existent priors (Johnson et al., 2019a). This was one of the main reasons why the CXR-PRO dataset was developed, as the reports in this dataset omit previous references to prevent hallucination. Fig. 9 shows the lack of publications that have used this data set to train their methods up to this date. The newly proposed methods should incorporate the CXR-PRO dataset in their training process to mitigate the issue of hallucination. Furthermore, novel evaluation methods must be developed taking this issue into account, penalizing the model when hallucinations occur in the evaluation.

5.4. Implications in real clinic practices

It is important to note that the inclusion of AI in medicine is not meant to replace, but rather to assist healthcare professionals in their daily work. In general, AI tools should be at the disposal of physicians for better medical outcomes, both for patients and healthcare professionals (Choy et al., 2018; Hardy and Harvey, 2020; Marcu and Marcu, 2021; Moor et al., 2023). Using AI capabilities, physicians can obtain valuable information, improve diagnostic accuracy, streamline workflows, and make more informed decisions.

Most AI-related work compares how a model performs compared to a physician. In fact, the precision of a model prediction is of great importance, but in real clinical practice, AI must be evaluated based on how it complements the skills and expertise of physicians with the objective of improving their performance (Langlotz, 2019). This idea also extrapolates to Medical Image Captioning and Diagnostic Captioning. Moor et al. (2023) identify several areas of application for what they define as General Medical Artificial Intelligence (GMA). Three areas are outlined in this work: bedside decision support, grounded radiology reports, and augmented procedures. The second and third areas are closely related to the radiology report task. They argue that these models should be able to pair text with interactive visualizations and consider the patient's clinical history. There are already studies that exploit the time differences between images from different visits of the same patients to assess the development of particular diseases (Bannur et al., 2023). Other works, such as the rAID.ologist framework (Amador-Domínguez et al., 2021), propose a case-based reasoning framework to assist physicians in the generation of medical reports. This framework, which is fully explainable and user-focused, provides formal corrections, references, and suggestions.

Ethics and privacy are two of the main practical challenges in integrating MIC and DC technologies into clinical workflows. In general, the use of generative models for medicine can introduce unintentional privacy breaches in the large data sets from which these models are trained (Li et al., 2023). Methods such as differential privacy, federated learning, and secure multiparty computation have been proposed in the literature to mitigate this problem (Koochi-Moghadam and Bae, 2023). Regarding ethics, the use of generative models trained from large datasets always involves the risk of bias in the captions generated (Schwarz et al., 2021). Methods proposed to mitigate bias include resampling, reweighting, and adversarial training (González-Sendino et al., 2023). A careful curation of the data to ensure diversity and a rigorous evaluation of the performance of the model in different demographic groups are essential to provide a viable solution (Koochi-Moghadam and Bae, 2023).

6. Conclusions and future work

The interest in AI-based generative text and image modes is slowly increasing by the scientific community, industry, and society. Image captioning is one of the key tasks in this area, as it combines two main areas of research within AI: computer vision and natural language processing. Natural Image Captioning (NIC) aims to extract key information from an image and generate a description in a human-like manner. Medical Image Captioning (MIC) is a subfield of this area which focuses on the description of medical images. Medical Image Captioning (MIC), Diagnostic Captioning (DC), and Medical Report Generation represent distinct but interconnected concepts in this domain. Although MIC refers to a more general concept of medical image descriptions, DC extends to diagnostic information about the medical interpretation of a patient's radiological image, although both terms are commonly interchanged in the literature. Automatic medical report generation includes DC as well as providing indications, recommendations, and temporal associations from previous visits (worsening or improvement).

This paper presents a survey of the existing literature and current developments and limitations in the field of image captioning, with a special focus on Medical Image Captioning (MIC), specifically Diagnostic Captioning (DC). The methodology followed to find and analyze the different NIC and DC resources is provided. The present survey provided a wide overview of the different existing NIC models, ranging from the early, retrieval, and template-based methods; to the more recent deep learning-based approaches. Most of the methods reviewed for NIC can also be applied for DC serving as a baseline for several relevant methods. This survey focuses especially on the applications of DC in the context of radiology, providing a broad overview of the most relevant datasets, existing evaluation methods, and existing models. In the context of datasets, their features and usage over time were discussed, highlighting the importance of the MIMIC-CXR dataset. The presented survey serves as a foundation for further research and development of improved captioning models, especially for medical image captioning.

In the case of DC models, it is worth mentioning that, while most of them are trained following a supervised learning approach, alternative learning strategies such as reinforcement learning are starting to emerge showing promising results. Attention-based models, compositional architectures, and hybrid methods that combine template- and retrieval-based models with generative approaches have also contributed significantly to MIC.

In the present survey, the shortcomings and advantages of existing DC approaches are also discussed. The main drawback of these models is their lack of explainability in the given captions, as they are unable to locate pathologies or anomalies in an image with precision. Dense captioning methods are capable of accurately locating elements in the image. However, explainability remains an issue in these methods. Another important limitation lies in the difference in perspective in NIC and DC. NIC aims to describe the full image, while DC focuses only on the relevant elements. This hinders the reusability and extension of NIC models for DC. Finally, hallucinations have also been shown to be a common issue in MIC models. Novel datasets, such as CXR-PRO, and the proposal of new evaluation methods that penalize hallucination have been proposed to mitigate this issue. A potential research avenue to address both identified limitations involves designing approaches capable of leveraging attention scores to: (1) generate explanations for the generated caption; and (2) discard irrelevant elements in the medical image. In this context, Beddiar et al. (2022a) propose an architecture to improve the explainability and trustworthiness of medical image diagnosis tasks based on these attention scores. The significant advantage of focusing on attention mechanisms is that most current Deep Learning models, including Large Language Models (LLMs), utilize them through the Transformers architecture. Hence, methodologies relying on attention scores can remain relatively independent of specific MIC and DC

methods, as long as these methods are grounded in the Transformer architecture. However, as highlighted by Zhao et al. (2023), who provides an overview of the methods to explain Transformer-based language models, the internal mechanisms of LLMs are still unclear and their lack of transparency is far from being solved.

One of the main lines of future work is the use of Vision-language models or Multimodal Large Language Models (M-LLMs). These models can interpret and generate not only text but also images. Since medicine is a multimodal discipline, these models have drastically reduced the complexity of using AI for medical purposes by healthcare professionals (Meskó, 2023). Yan et al. (2023) analyze GPT-4 with Vision (GPT-4V) on the task of Visual Question Answering (VQA). The model was evaluated to answer questions paired with images using pathology and radiology datasets. The authors conclude that the use of these pre-trained models is not recommended for real-world diagnostics because they are unreliable and offer less accuracy than rival alternatives. Sorin et al. (2023) reached similar conclusions when using GPT-4V in ophthalmology. The model achieved a correct diagnosis in 67.5% cases when a clinical context was provided. Although the authors concluded that this M-LLM is not suitable for clinical applications in ophthalmology, they acknowledge their ability to simultaneously analyze and integrate visual and textual data. Horiuchi et al. (2024) evaluated the diagnostic performance of GPT-4 based ChatGPT in neuroradiology by providing the model with the patient's medical history and imaging findings. The diagnostic accuracy rate obtained was 50% (50/100 cases). It is important to note that M-LLMs are a very immature technology: the first of these models, GPT-4, was introduced in March 2023. The use of M-LLMs has the potential to process vast textual and multimodal data to provide insights into the relationships between radiological images, clinical reports, and patient outcomes (Koohi-Moghadam and Bae, 2023).

Addressing the limitations exposed in this study mostly involves having more and better data to train (or to fine-tune) Deep Learning models. With this in mind, two main directions in future research are essential. First, future data and models should include temporal information, such as the clinical history of patients. In addition, different types of nonvisual information, such as laboratory results or relevant clinical findings. Developing methods to effectively integrate temporal information and heterogeneous input would improve the quality and usefulness of the final reports. Secondly, since datasets for DC are mostly limited to web-scraped image-text pairs, an essential research direction is the development of larger and more diverse datasets that should be multimodal and multicenter.

CRedit authorship contribution statement

Gabriel Reale-Nosei: Writing – original draft, Methodology, Investigation, Conceptualization. **Elvira Amador-Domínguez:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Conceptualization. **Emilio Serrano:** Writing – review & editing, Supervision, Conceptualization.

Declaration of competing interest

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

Data availability

No data was used for the research described in the article.

Acknowledgments

This research work has been funded by the KnowledgeSpaces Project (Grant PID2020-118274RB-I00 funded by MCIN/AEI/10.13039/501100011033).

References

- Amador-Domínguez, E., Serrano, E., Manrique, D., Bajo, J., 2021. A case-based reasoning model powered by deep learning for radiology report recommendation. Publisher: International Journal of Interactive Multimedia and Artificial Intelligence
- Ambati, R., Dudyala, C.R., 2018. A sequence-to-sequence model approach for imageclef 2018 medical domain visual question answering. In: 2018 15th IEEE India Council International Conference (INDICON). IEEE, pp. 1–6.
- Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L., 2018. Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
- Ayesha, H., Iqbal, S., Tariq, M., Abrar, M., Sanaullah, M., Abbas, I., Rehman, A., Niazi, M.F.K., Hussain, S., 2021. Automatic medical image interpretation: State of the art and future directions. Pattern Recognit. 114, 107856. <http://dx.doi.org/10.1016/j.patcog.2021.107856>, URL: <https://www.sciencedirect.com/science/article/pii/S0031320321000431>.
- Bai, S., An, S., 2018. A survey on automatic image caption generation. Neurocomputing 311, 291–304. <http://dx.doi.org/10.1016/j.neucom.2018.05.080>, URL: <https://www.sciencedirect.com/science/article/pii/S0925231218306659>.
- Banerjee, S., Lavie, A., 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Association for Computational Linguistics, Ann Arbor, Michigan, pp. 65–72, URL: <https://aclanthology.org/W05-0909>.
- Bannur, S., Hyland, S., Liu, Q., Perez-Garcia, F., Ilse, M., Castro, D.C., Boecking, B., Sharma, H., Bouzid, K., Thieme, A., et al., 2023. Learning to exploit temporal structure for biomedical vision-language processing.
- Beddier, D.R., Oussalah, M., Seppänen, T., 2022a. Explainability for medical image captioning. In: Eleventh International Conference on Image Processing Theory, Tools and Applications, IPTA 2022, Salzburg, Austria, April 19–22, 2022. IEEE, pp. 1–6. <http://dx.doi.org/10.1109/IPTA54936.2022.9784146>.
- Beddier, D.R., Oussalah, M., Seppänen, T., 2022b. Automatic captioning for medical imaging (MIC): a rapid review of literature. Artif. Intell. Rev. <http://dx.doi.org/10.1007/s10462-022-10270-w>.
- Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A., Plank, B., 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. J. Artificial Intelligence Res. 55, 409–442. <http://dx.doi.org/10.1613/jair.4900>, URL: <https://www.jair.org/index.php/jair/article/view/10985>.
- Boecking, B., Usuyama, N., Bannur, S., Coelho de Castro, D., Schwaighofer, A., Hyland, S.L., Wetscherek, M.T., Naumann, T., Nori, A.V., Alvarez Valle, J., Poon, H., Oraz, O., 2022. MS-CXR: Making the most of text semantics to improve biomedical vision-language processing. URL: <http://dx.doi.org/10.13026/b90j-vb87>.
- Bustos, A., Pertusa, A., Salinas, J.-M., de la Iglesia-Vayá, M., 2020. Padchest: A large chest x-ray image dataset with multi-label annotated reports. Med. Image Anal. 66, 101797, Publisher: Elsevier.
- Charalampakos, F., Karatzas, V., Kougia, V., Pavlopoulos, J., Androutsopoulos, I., 2021. AUEB NLP group at ImageCLEFmed caption tasks 2021. In: CLEF (Working Notes). pp. 1184–1200.
- Chelaramani, S., Gupta, M., Agarwal, V., Gupta, P., Habash, R., 2020. Multi-task learning for fine-grained eye disease prediction. In: Pattern Recognition: 5th Asian Conference, ACPR 2019, Auckland, New Zealand, November 26–29, 2019, Revised Selected Papers, Part II. Springer, pp. 734–749.
- Chen, Z., Song, Y., Chang, T.-H., Wan, X., 2020. Generating radiology reports via memory-driven transformer. arXiv preprint [arXiv:2010.16056](https://arxiv.org/abs/2010.16056).
- Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., Chua, T.-S., 2017. SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
- Choy, G., Khalilzadeh, O., Michalski, M., Do, S., Samir, A.E., Panykh, O.S., Geis, J.R., Pandharipande, P.V., Brink, J.A., Dreyer, K.J., 2018. Current applications and future impact of machine learning in radiology. Radiology 288 (2), 318–328. <http://dx.doi.org/10.1148/radiol.2018171820>.
- Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R., 2020. Meshed-memory transformer for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10578–10587.
- Dai, B., Fidler, S., Urtasun, R., Lin, D., 2017. Towards diverse and natural image descriptions via a conditional GAN. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 2989–2998.
- Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooashan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J., 2016. Preparing a collection of radiology examinations for distribution and retrieval. J. Am. Med. Inform. Assoc. 23 (2), 304–310, Publisher: Oxford University Press.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint [arXiv:2010.11929](https://arxiv.org/abs/2010.11929).

- Fang, H., Gupta, S., Iandola, F., Srivastava, R.K., Deng, L., Dollar, P., Gao, J., He, X., Mitchell, M., Platt, J.C., Lawrence Zitnick, C., Zweig, G., 2015. From captions to visual concepts and back. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Farhadi, A., Hejrati, M., Sadeghi, M.A., Young, P., Rashtchian, C., Hockenmaier, J., Forsyth, D., 2010. Every picture tells a story: Generating sentences from images. In: *Daniilidis, K., Maragos, P., Paragios, N. (Eds.), Computer Vision – ECCV 2010*. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 15–29.
- Gale, W., Oakden-Rayner, L., Carneiro, G., Palmer, L.J., Bradley, A.P., 2019. Producing radiologist-quality reports for interpretable deep learning. In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. pp. 1275–1279. <http://dx.doi.org/10.1109/ISBI.2019.8759236>.
- Gong, Y., Wang, L., Hodosh, M., Hockenmaier, J., Lazebnik, S., 2014. Improving image-sentence embeddings using large weakly annotated photo collections. In: *Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (Eds.), Computer Vision – ECCV 2014*. Springer International Publishing, Cham, pp. 529–545.
- González-Sendino, R., Serrano, E., Bajo, J., Novais, P., 2023. A review of bias and fairness in artificial intelligence. *Int. J. Interact. Multimedia Artif. Intell.* 1–13. <http://dx.doi.org/10.9781/ijimai.2023.11.001>, in press.
- Graham, Y., 2015. Re-evaluating automatic summarization with BLEU and 192 shades of ROUGE. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pp. 128–137. <http://dx.doi.org/10.18653/v1/D15-1013>, URL: <https://aclanthology.org/D15-1013>.
- Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H., 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthc. (HEALTH)* 3 (1), 1–23, Publisher: ACM New York, NY.
- Guo, L., Liu, J., Zhu, X., Yao, P., Lu, S., Lu, H., 2020. Normalized and geometry-aware self-attention network for image captioning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 10327–10336.
- Han, Z., Wei, B., Leung, S., Chung, J., Li, S., 2018. Towards automatic report generation in spine radiology using weakly supervised framework. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16–20, 2018, Proceedings, Part IV 11*. Springer, pp. 185–193.
- Han, Z., Wei, B., Xi, X., Chen, B., Yin, Y., Li, S., 2021. Unifying neural learning and symbolic reasoning for spinal medical report generation. *Med. Image Anal.* 67, 101872, Publisher: Elsevier.
- Hardy, M., Harvey, H., 2020. Artificial intelligence in diagnostic imaging: impact on the radiography profession. *Br. J. Radiol.* 93 (1108), 20190840, Publisher: The British Institute of Radiology.
- Harzig, P., Einfalt, M., Lienhart, R., 2019. Automatic disease detection and report generation for gastrointestinal tract examination. In: *Proceedings of the 27th ACM International Conference on Multimedia*. MM '19, Association for Computing Machinery, New York, NY, USA, pp. 2573–2577. <http://dx.doi.org/10.1145/3343031.3356066>, event-place: Nice, France.
- Hodosh, M., Young, P., Hockenmaier, J., 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artificial Intelligence Res.* 47, 853–899. <http://dx.doi.org/10.1613/jair.3994>, URL: <https://jair.org/index.php/jair/view/10833>.
- Horiuchi, D., Tatekawa, H., Shimono, T., Walston, S.L., Takita, H., Matsushita, S., Oura, T., Mitsuyama, Y., Miki, Y., Ueda, D., 2024. Accuracy of ChatGPT generated diagnosis from patient's medical history and imaging findings in neuroradiology cases. *Neuroradiology* 66 (1), 73–79. <http://dx.doi.org/10.1007/s00234-023-03252-4>, Epub 2023 Nov 23.
- Hossain, M.Z., Sohel, F., Shiratuddin, M.F., Laga, H., 2019. A comprehensive survey of deep learning for image captioning. *ACM Comput. Surv.* 51 (6), 118:1–118:36. <http://dx.doi.org/10.1145/3295748>.
- Hou, B., Kaissis, G., Summers, R.M., Kainz, B., 2021. Ratchet: Medical transformer for chest x-ray diagnosis and reporting. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VII 24*. Springer, pp. 293–303.
- Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 4700–4708.
- Ionescu, B., Müller, H., Drăgulescu, A.M., Popescu, A., Idrissi-Yaghir, A., García Seco de Herrera, A., Andrei, A., Stan, A., Storás, A.M., Abacha, A.B., Friedrich, C.M., Ioannidis, G., Adams, G., Schäfer, H., Manguinhas, H., Filipovich, I., Coman, I., Deshayes, J., Schöler, J., Rückert, J., Ștefan, L.-D., Bloch, L., Yetisgen, M., Riegler, M.A., Dogariu, M., Constantin, M.G., Snider, N., Papachrysos, N., Halvorsen, P., Brüngel, R., Kozlovski, S., Hicks, S., de Lange, T., Thambawita, V., Kovalev, V., Yim, W.-W., 2023. ImageCLEF 2023 Highlight: Multimedia retrieval in medical, social media and content recommendation applications. In: *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part III*. Springer-Verlag, Berlin, Heidelberg, pp. 557–567. http://dx.doi.org/10.1007/978-3-031-28241-6_64, event-place: Dublin, Ireland.
- Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghighi, B., Ball, R., Shpanskaya, K., et al., 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33, pp. 590–597, Issue: 01.
- Jain, S., Agrawal, A., Saporta, A., Truong, S.Q., Duong, D.N., Bui, T., Chambon, P., Zhang, Y., Lungren, M.P., Ng, A.Y., et al., 2021. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*.
- Jing, B., Wang, Z., Xing, E., 2020. Show, describe and conclude: On exploiting the structure information of chest x-ray reports. *arXiv preprint arXiv:2004.12274*.
- Jing, B., Xie, P., Xing, E., 2018. On the automatic generation of medical imaging reports. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, pp. 2577–2586. <http://dx.doi.org/10.18653/v1/P18-1240>, URL: <https://aclanthology.org/P18-1240>.
- Johnson, J., Karpathy, A., Fei-Fei, L., 2016. Denscap: Fully convolutional localization networks for dense captioning. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 4565–4574.
- Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.-y., Mark, R.G., Horng, S., 2019a. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci. Data* 6 (1), 317, Publisher: Nature Publishing Group UK London.
- Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.-y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S., 2019b. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.
- Karatzas, B., Pavlopoulos, J., Kougia, V., Androutsopoulos, I., 2020. AUEB NLP group at ImageCLEFmed caption task 2020. In: *CLEF 2020 Working Notes*, Thessaloniki, Greece, September 22–25, 2020.
- Karpathy, A., Fei-Fei, L., 2015. Deep visual-semantic alignments for generating image descriptions. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Ke, L., Pei, W., Li, R., Shen, X., Tai, Y.-W., 2019. Reflective decoding network for image captioning. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
- Kipf, T.N., Welling, M., 2017. Semi-supervised classification with graph convolutional networks. *eprint: 1609.02907*.
- Kisilev, P., Walach, E., Barkan, E., Ophir, B., Alpert, S., Hashoul, S.Y., 2015. From medical image to automatic medical report generation. *IBM J. Res. Dev.* 59 (2/3), 2:1–2:7. <http://dx.doi.org/10.1147/JRD.2015.2393193>.
- Koohi-Moghadam, M., Bae, K.T., 2023. Generative AI in medical imaging: Applications, challenges, and ethics. *J. Med. Syst.* 47 (1), 94. <http://dx.doi.org/10.1007/s10916-023-01987-4>.
- Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D.A., et al., 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.* 123, 32–73, Publisher: Springer.
- Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks. In: *Pereira, F., Burges, C.J., Bottou, L., Weinberger, K.Q. (Eds.), In: Advances in Neural Information Processing Systems*, vol. 25, Curran Associates, Inc., URL: <https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf>.
- Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., Berg, A.C., Berg, T.L., 2013. BabyTalk: Understanding and generating simple image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.* 35 (12), 2891–2903. <http://dx.doi.org/10.1109/TPAMI.2012.162>.
- Kumar, A., Goel, S., 2017. A survey of evolution of image captioning techniques. *Int. J. Hybrid Intell. Syst.* 14 (3), 123–139. <http://dx.doi.org/10.3233/HIS-170246>, Place: NLD Publisher: IOS Press.
- Langlotz, C.P., 2019. Will artificial intelligence replace radiologists?. *Issue: 3 Pages: e190058 Publication Title: Radiology: Artificial Intelligence Volume: 1*.
- Li, X., Cao, R., Zhu, D., 2019c. Vispi: Automatic visual perception and interpretation of chest x-rays. *arXiv preprint arXiv:1906.05190*.
- Li, S., Kulkarni, G., Berg, T.L., Berg, A.C., Choi, Y., 2011. Composing simple image descriptions using web-scale N-grams. In: *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Portland, Oregon, USA, pp. 220–228, URL: <https://aclanthology.org/W11-0326>.
- Li, Y., Liang, X., Hu, Z., Xing, E.P., 2018. Hybrid retrieval-generation reinforced agent for medical image report generation. In: *Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (Eds.), In: Advances in Neural Information Processing Systems*, vol. 31, Curran Associates, Inc., URL: https://proceedings.neurips.cc/paper_files/paper/2018/file/e07413354875b01a996dc560274708e-Paper.pdf.
- Li, C.Y., Liang, X., Hu, Z., Xing, E.P., 2019a. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33, pp. 6666–6673. <http://dx.doi.org/10.1609/aaai.v33i01.33016666>, URL: <https://ojs.aaai.org/index.php/AAAI/article/view/46637>, Number: 01.

- Li, H., Moon, J.T., Purkayastha, S., Celi, L.A., Trivedi, H., Gichoya, J.W., 2023. Ethics of large language models in medicine and medical research. *Lancet Digit. Health* 5 (6), e333–e335. [http://dx.doi.org/10.1016/S2589-7500\(23\)00083-3](http://dx.doi.org/10.1016/S2589-7500(23)00083-3), URL: <https://www.sciencedirect.com/science/article/pii/S2589750023000833>.
- Li, J., Yao, P., Guo, L., Zhang, W., 2019b. Boosted transformer for image captioning. *Appl. Sci.* 9 (16), 3260, Publisher: MDPI.
- Lin, C.-Y., 2004. ROUGE: A package for automatic evaluation of summaries. In: *Text Summarization Branches Out. Association for Computational Linguistics*, Barcelona, Spain, pp. 74–81, URL: <https://aclanthology.org/W04-1013>.
- Liu, G., Hsu, T.-M.H., McDermott, M., Boag, W., Weng, W.-H., Szolovits, P., Ghassemi, M., 2019. Clinically accurate chest X-ray report generation. In: Doshi-Velez, F., Fackler, J., Jung, K., Kale, D., Ranganath, R., Wallace, B., Wiens, J. (Eds.), *Proceedings of the 4th Machine Learning for Healthcare Conference*. In: *Proceedings of Machine Learning Research*, vol. 106, PMLR, pp. 249–269, URL: <https://proceedings.mlr.press/v106/liu19a.html>.
- Liu, X., Xu, Q., Wang, N., 2018. A survey on deep neural network-based image captioning. *Vis. Comput.* 35 (3), 445–470. <http://dx.doi.org/10.1007/s00371-018-1566-y>.
- Lu, J., Xiong, C., Parikh, D., Socher, R., 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Marcu, L.G., Marcu, D., 2021. Points of view on artificial intelligence in medical imaging—One good, one bad, one fuzzy. *Health Technol.* 11, 17–22, Publisher: Springer.
- Meskó, B., 2023. The impact of multimodal large language models on health care's future. *J. Med. Internet Res.* 25 (1), e52865. <http://dx.doi.org/10.2196/52865>.
- Messina, P., Pino, P., Parra, D., Soto, A., Besa, C., Uribe, S., Andía, M., Tejos, C., Prieto, C., Capurro, C., 2022. A survey on deep learning and explainability for automatic report generation from medical images. *ACM Comput. Surv.* 54 (10s), 203:1–203:40. <http://dx.doi.org/10.1145/3522747>, URL: <https://dl.acm.org/doi/10.1145/3522747>.
- Miura, Y., Zhang, Y., Tsai, E.B., Langlotz, C.P., Jurafsky, D., 2020. Improving factual completeness and consistency of image-to-text radiology report generation. *arXiv preprint arXiv:2010.10042*.
- Mohsan, M.M., Akram, M.U., Rasool, G., Alghamdi, N.S., Baqai, M.A.A., Abbas, M., 2022. Vision transformer and language model based radiology report generation. *IEEE Access* 11, 1814–1824, Publisher: IEEE.
- Monshi, M.M.A., Poon, J., Chung, V., 2020. Deep learning in generating radiology reports: A survey. *Artif. Intell. Med.* 106, 101878. <http://dx.doi.org/10.1016/j.artmed.2020.101878>, URL: <https://www.sciencedirect.com/science/article/pii/S0933365719302635>.
- Moor, M., Banerjee, O., Abad, Z.S.H., Krumholz, H.M., Leskovec, J., Topol, E.J., Rajpurkar, P., 2023. Foundation models for generalist medical artificial intelligence. *Nature* 616 (7956), 259–265, Publisher: Nature Publishing Group UK London.
- Mork, J.G., Jimeno-Yepes, A., Aronson, A.R., 2013. The NLM medical text indexer system for indexing biomedical literature. In: *BioASQ@CLEF*.
- Nazir, S., Dickson, D.M., Akram, M.U., 2023. Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks. *Comput. Biol. Med.* 156, 106668. <http://dx.doi.org/10.1016/j.combiomed.2023.106668>, URL: <https://www.sciencedirect.com/science/article/pii/S0010482523001336>.
- Nicolson, A., Dowling, J., Koopman, B., 2021. AEHRCSIRO at ImageCLEFmed caption 2021. In: *CLEF (Working Notes)*. pp. 1317–1328.
- Ordóñez, V., Kulkarni, G., Berg, T., 2011. Im2Text: Describing images using 1 million captioned photographs. In: Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., Weinberger, K.Q. (Eds.), *Advances in Neural Information Processing Systems*, vol. 24, Curran Associates, Inc., URL: <https://proceedings.neurips.cc/paper/2011/file/5dd9db5e033da9c6fb5ba83c7a7eb9a-Paper.pdf>.
- Ouyang, X., Karanam, S., Wu, Z., Chen, T., Huo, J., Zhou, X.S., Wang, Q., Cheng, J.-Z., 2020. Learning hierarchical attention for weakly-supervised chest X-ray abnormality localization and diagnosis. *IEEE Trans. Med. Imaging* 40 (10), 2698–2710, Publisher: IEEE.
- Pan, Y., Yao, T., Li, Y., Mei, T., 2020. X-linear attention networks for image captioning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 10971–10980.
- Papineni, K., Roukos, S., Ward, T., Zhu, W.-J., 2002. Bleu: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pp. 311–318. <http://dx.doi.org/10.3115/1073083.1073135>, URL: <https://aclanthology.org/P02-1040>.
- Pavlopoulos, J., Kougia, V., Androutsopoulos, I., 2019. A Survey on Biomedical Image Captioning. *Association for Computational Linguistics*, Minneapolis, Minnesota, pp. 26–36. <http://dx.doi.org/10.18653/v1/W19-1803>, URL: <https://aclanthology.org/W19-1803>.
- Pavlopoulos, J., Kougia, V., Androutsopoulos, I., Papamichail, D., 2022. Diagnostic captioning: a survey. *Knowl. Inf. Syst.* 64 (7), 1691–1722. <http://dx.doi.org/10.1007/s10115-022-01684-7>.
- Pelka, O., Friedrich, C.M., et al., 2017. Keyword generation for biomedical image retrieval with recurrent neural networks. In: *CLEF (Working Notes)*.
- Pelka, O., Koitka, S., Rückert, J., Nensa, F., Friedrich, C.M., 2018. Radiology objects in context (ROCO): a multimodal image dataset. In: *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3*. Springer, pp. 180–189.
- Pino, P., Parra, D., Besa, C., Lagos, C., 2021. Clinically correct report generation from chest X-rays using templates. In: *Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 12*. Springer, pp. 654–663.
- Qin, Y., Du, J., Zhang, Y., Lu, H., 2019. Look back and predict forward in image captioning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Ramesh, V., Chi, N., Rajpurkar, P., 2019. CXR-PRO: MIMIC-CXR with prior references omitted (version 1.0.0). Publisher: Nature Publishing Group UK London.
- Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (Eds.), *Advances in Neural Information Processing Systems*, vol. 28, Curran Associates, Inc., URL: <https://proceedings.neurips.cc/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf>.
- Ren, Z., Wang, X., Zhang, N., Lv, X., Li, L.-J., 2017. Deep reinforcement learning-based image captioning with embedding reward. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 1151–1159. <http://dx.doi.org/10.1109/CVPR.2017.128>.
- Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V., 2017. Self-critical sequence training for image captioning. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 1179–1195. <http://dx.doi.org/10.1109/CVPR.2017.131>.
- Rückert, J., Ben Abacha, A., G. Seco de Herrera, A., Bloch, L., Brüngel, R., Idrissi-Yaghir, A., Schäfer, H., Müller, H., Friedrich, C.M., 2023. Overview of ImageCLEFmedical 2023 – caption prediction and concept detection. In: *CLEF2023 Working Notes*. In: *CEUR Workshop Proceedings*, CEUR-WS.org, Thessaloniki, Greece.
- Sammani, F., Melas-Kyriazi, L., 2020. Show, edit and tell: a framework for editing image captions. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 4808–4816.
- Schwarz, K., Liao, Y., Geiger, A., 2021. On the frequency bias of generative models. In: Ranzato, M., Beygelzimer, A., Dauphin, Y.N., Liang, P., Vaughan, J.W. (Eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6–14, 2021, Virtual*. pp. 18126–18136, URL: <https://proceedings.neurips.cc/paper/2021/hash/96bf57c6ff19504ff145e2a32991ea96-Abstract.html>.
- Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE International Conference on Computer Vision*. pp. 618–626.
- Shamshad, F., Khan, S., Zamir, S.W., Khan, M.H., Hayat, M., Khan, F.S., Fu, H., 2023. Transformers in medical imaging: A survey. *Med. Image Anal.* 102802. <http://dx.doi.org/10.1016/j.media.2023.102802>, URL: <https://www.sciencedirect.com/science/article/pii/S1361841523000634>.
- Shao, Z., Han, J., Mamerides, D., Debattista, K., 2022. Region-object relation-aware dense captioning via transformer. *IEEE Trans. Neural Netw. Learn. Syst.* 1–12. <http://dx.doi.org/10.1109/TNNLS.2022.3152990>.
- Shetty, R., Rohrbach, M., Anne Hendricks, L., Fritz, M., Schiele, B., 2017. Speaking the same language: Matching machine to human captions by adversarial training. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.
- Singh, S., Karimi, S., Ho-Shon, K., Hamey, L., 2019. From chest x-rays to radiology reports: a multimodal machine learning approach. In: *2019 Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, pp. 1–8.
- Sorin, V., Kapelushnik, N., Hecht, I., Zlot, O., Glicksberg, B., Bufman, H., Barash, Y., Nadkarni, G., Klang, E., 2023. GPT-4 multimodal analysis on ophthalmology clinical cases including text and images. *medRxiv* <http://dx.doi.org/10.1101/2023.11.24.23298953>.
- Sun, C., Gan, C., Nevatia, R., 2015. Automatic concept discovery from parallel text and visual corpora. In: *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*. ICCV '15, IEEE Computer Society, USA, pp. 2596–2604. <http://dx.doi.org/10.1109/ICCV.2015.298>.
- Sun, L., Wang, W., Li, J., Lin, J., 2019. Study on medical image report generation based on improved encoding-decoding method. In: *Intelligent Computing Theories and Application: 15th International Conference, ICIC 2019, Nanchang, China, August 3–6, 2019, Proceedings, Part I*. Springer, pp. 686–696.
- Sutskever, I., Vinyals, O., Le, Q.V., 2014. Sequence to sequence learning with neural networks. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K.Q. (Eds.), *Advances in Neural Information Processing Systems*, vol. 27, Curran Associates, Inc., URL: <https://proceedings.neurips.cc/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf>.
- Syeda-Mahmood, T., Wong, K.C., Gur, Y., Wu, J.T., Jadhav, A., Kashyap, S., Karargiris, A., Pillai, A., Sharma, A., Syed, A.B., et al., 2020. Chest x-ray report generation through fine-grained label learning. In: *Medical Image Computing and Computer Assisted Intervention—MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II*. Springer, pp. 561–571.

- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Tsuned, R., Asakawa, T., Aono, M., 2021. Kdelab at imageclef 2021: Medical caption prediction with effective data pre-processing and deep learning. In: *CLEF (Working Notes)*. pp. 1365–1374.
- Ushiku, Y., Yamaguchi, M., Mukuta, Y., Harada, T., 2015. Common subspace for model and similarity: Phrase learning for caption generation from images. In: *2015 IEEE International Conference on Computer Vision (ICCV)*. pp. 2668–2676. <http://dx.doi.org/10.1109/ICCV.2015.306>.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., URL: <https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>.
- Vedantam, R., Zitnick, C.L., Parikh, D., 2015. CIDEr: Consensus-based image description evaluation. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 4566–4575. <http://dx.doi.org/10.1109/CVPR.2015.7299087>.
- Vinyals, O., Toshev, A., Bengio, S., Erhan, D., 2015. Show and tell: A neural image caption generator. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Wang, L., Bai, Z., Zhang, Y., Lu, H., 2020. Show, recall, and tell: Image captioning with recall mechanism. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34, pp. 12176–12183, Issue: 07.
- Wang, Z., Liu, L., Wang, L., Zhou, L., 2023. METransformer: Radiology report generation by transformer with multiple learnable expert tokens. *arXiv preprint arXiv:2304.02211*.
- Wang, X., Peng, Y., Lu, L., Lu, Z., Summers, R.M., 2018a. TieNet: Text-image embedding network for common thorax disease classification and reporting in chest X-rays. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 9049–9058. <http://dx.doi.org/10.1109/CVPR.2018.00943>.
- Wang, X., Zhang, Y., Guo, Z., Li, J., 2018b. ImageSem at imageclef 2018 caption task: Image retrieval and transfer learning. In: *Conference and Labs of the Evaluation Forum*.
- Wu, J.T., Agu, N.N., Lourentzou, I., Sharma, A., Paguio, J.A., Yao, J.S., Dee, E.C., Mitchell, W., Kashyap, S., Giovannini, A., et al., 2021. Chest ImaGenome dataset for clinical reasoning. *arXiv preprint arXiv:2108.00316*.
- Xie, X., Xiong, Y., Yu, P.S., Li, K., Zhang, S., Zhu, Y., 2019. Attention-based abnormal-aware fusion network for radiology report generation. In: *Database Systems for Advanced Applications: DASFAA 2019 International Workshops: BDMS, BDQM, and GDMA*, Chiang Mai, Thailand, April 22–25, 2019, *Proceedings 24*. Springer, pp. 448–452.
- Xiong, Y., Du, B., Yan, P., 2019. Reinforced transformer for medical image captioning. In: *Machine Learning in Medical Imaging: 10th International Workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 10*. Springer, pp. 673–680.
- Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y., 2015. Show, attend and tell: Neural image caption generation with visual attention. In: Bach, F., Blei, D. (Eds.), *Proceedings of the 32nd International Conference on Machine Learning*. In: *Proceedings of Machine Learning Research*, Vol. 37, PMLR, Lille, France, pp. 2048–2057, URL: <https://proceedings.mlr.press/v37/xuc15.html>.
- Xu, J., Liu, W., Liu, C., Wang, Y., Chi, Y., Xie, X., Hua, X.-S., 2019. Concept detection based on multi-label classification and image captioning approach-DAMO at imageclef 2019. In: *CLEF (Working Notes)*.
- Xue, Y., Huang, X., 2019. Improved disease classification in chest x-rays with transferred features from report generation. In: *Information Processing in Medical Imaging: 26th International Conference, IPMI 2019, Hong Kong, China, June 2–7, 2019, Proceedings 26*. Springer, pp. 125–138.
- Yan, Z., Zhang, K., Zhou, R., He, L., Li, X., Sun, L., 2023. Multimodal ChatGPT for medical applications: an experimental study of GPT-4V. <http://dx.doi.org/10.48550/ARXIV.2310.19061>, *CoRR abs/2310.19061* *arXiv:2310.19061*.
- Yang, L., Tang, K., Yang, J., Li, L.-J., 2017. Dense captioning with joint inference and visual context. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Yang, Y., Teo, C., Daumé III, H., Aloimonos, Y., 2011. Corpus-guided sentence generation of natural images. In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Edinburgh, Scotland, UK., pp. 444–454, URL: <https://aclanthology.org/D11-1041>.
- Yao, T., Pan, Y., Li, Y., Mei, T., 2019. Hierarchy parsing for image captioning. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 2621–2629.
- Yao, T., Pan, Y., Li, Y., Qiu, Z., Mei, T., 2017. Boosting image captioning with attributes. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. pp. 4904–4912. <http://dx.doi.org/10.1109/ICCV.2017.524>.
- Yin, C., Qian, B., Wei, J., Li, X., Zhang, X., Li, Y., Zheng, Q., 2019a. Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network. In: *2019 IEEE International Conference on Data Mining (ICDM)*. pp. 728–737. <http://dx.doi.org/10.1109/ICDM.2019.00083>.
- Yin, G., Sheng, L., Liu, B., Yu, N., Wang, X., Shao, J., 2019b. Context and attribute grounded dense captioning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 6241–6250.
- You, Q., Jin, H., Wang, Z., Fang, C., Luo, J., 2016. Image captioning with semantic attention. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Yuan, J., Liao, H., Luo, R., Luo, J., 2019. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*. Springer, pp. 721–729.
- Zeng, X.-H., Liu, B.-G., Zhou, M., 2018. Understanding and generating ultrasound image description. *J. Comput. Sci. Tech.* 33, 1086–1100, Publisher: Springer.
- Zeng, X., Wen, L., Liu, B., Qi, X., 2020a. Deep learning for ultrasound image caption generation based on object detection. *Neurocomputing* 392, 132–141. <http://dx.doi.org/10.1016/j.neucom.2018.11.114>, URL: <https://www.sciencedirect.com/science/article/pii/S0925231219304825>.
- Zeng, X., Wen, L., Xu, Y., Ji, C., 2020b. Generating diagnostic report for medical image by high-middle-level visual information incorporation on double deep learning models. *Comput. Methods Programs Biomed.* 197, 105700. <http://dx.doi.org/10.1016/j.cmpb.2020.105700>, URL: <https://www.sciencedirect.com/science/article/pii/S0169260720315339>.
- Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y., 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.
- Zhang, S., Metaxas, D., 2024. On the challenges and perspectives of foundation models for medical image analysis. *Med. Image Anal.* 91, 102996. <http://dx.doi.org/10.1016/j.media.2023.102996>, URL: <https://www.sciencedirect.com/science/article/pii/S1361841523002566>.
- Zhang, L., Sung, F., Liu, F., Xiang, T., Gong, S., Yang, Y., Hospedales, T.M., 2017a. Actor-critic sequence training for image captioning. *eprint: 1706.09601*.
- Zhang, Z., Xie, Y., Xing, F., McGough, M., Yang, L., 2017b. MDNet: A semantically and visually interpretable medical image diagnosis network. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 3549–3557. <http://dx.doi.org/10.1109/CVPR.2017.378>.
- Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., Du, M., 2023. Explainability for large language models: A survey. <http://dx.doi.org/10.48550/ARXIV.2309.01029>, *CoRR arXiv:2309.01029* *arXiv:2309.01029*.
- Zhou, H.-Y., Chen, X., Zhang, Y., Luo, R., Wang, L., Yu, Y., 2022. Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports. *Nat. Mach. Intell.* 4 (1), 32–40, Publisher: Nature Publishing Group UK London.
- Zohourianshahzadi, Z., Kalita, J.K., 2022. Neural attention for image captioning: review of outstanding methods. *Artif. Intell. Rev.* 55 (5), 3833–3862. <http://dx.doi.org/10.1007/s10462-021-10092-2>.