
DIAGNOSTIC CAPTIONING: A SURVEY

John Pavlopoulos^{1,*}, Vasiliki Kougia^{1,2}, Ion Androutsopoulos², and Dimitris Papamichail³

¹Department of Computer and Systems Sciences, Stockholm University, Borgarfjordsgatan 12, 164
55 Kista, Sweden

²Department of Informatics, Athens University of Economics and Business, Patission 76, GR-104
34, Athens, Greece

³Nuclear Medicine Department, Medical Diagnostic Center Kosmoiatriki, Patission 237, GR-112
54, Athens, Greece

*Corresponding author. Email: annis@aeub.gr

ABSTRACT

Diagnostic Captioning (DC) concerns the automatic generation of a diagnostic text from a set of medical images of a patient collected during an examination. DC can assist inexperienced physicians, reducing clinical errors. It can also help experienced physicians produce diagnostic reports faster. Following the advances of deep learning, especially in generic image captioning, DC has recently attracted more attention, leading to several systems and datasets. This article is an extensive overview of DC. It presents relevant datasets, evaluation measures, and up to date systems. It also highlights shortcomings that hinder DC's progress and proposes future directions.

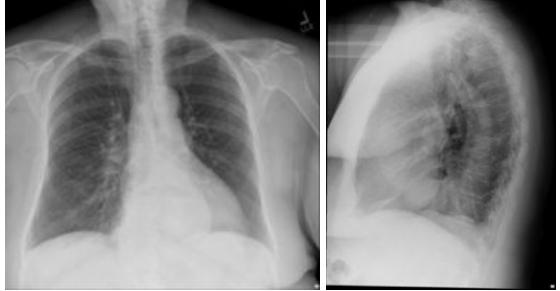
1 Introduction

Medical Imaging is concerned with forming visual representations of the anatomy or a function of a human body using a variety of imaging modalities (e.g., X-rays, CT, MRI) (Suetens, 2009; Aerts et al., 2014). It is estimated that approximately one billion medical imaging examinations are performed worldwide annually (Krupinski, 2010), and the number of medical imaging examinations per year continues to rise in developed countries (Brady, 2017). More sophisticated medical imaging systems lead to more images per examination and radiologist (Chokshi et al., 2015). The total workload has increased by 26% from 1998 to 2010 (Chokshi et al., 2015), and radiologists must now interpret more images during work time compared to similar examinations performed 10–20 years ago. Radiologists need to consider the examination's medical images, patient history, previous examinations, consult recent bibliography, and prepare a medical report. An increased workload in this demanding task increases the likelihood of medical errors (e.g., when radiologists are tired or pressured). These errors are not rare (Berlin, 2007) and they will also be present in the medical report, which is what referring physicians (who ordered the examination) mostly consider. Consequently, tools that would help radiologists produce higher quality reports (e.g., without missing important findings or reporting wrong findings when they are inexperienced) in less time (e.g., by providing them with a draft report) could have a significant impact.

Diagnostic Captioning (DC) systems encode the medical images of a patient's examination (or 'study') and generate a full or partial draft of the report. Fig. 1(b), for example, shows only the 'FINDINGS' section of the report. Although we are not aware of any clearly articulated description of the exact goals of research on DC systems, the main goals seem to be to (a) increase the *throughput* of medical imaging departments (Hosny et al., 2018), (b) reduce *medical errors* (Fazal et al., 2018), and (c) reduce the *cost* of medical imaging examinations (Liew, 2018). Advances in image to text generation (Bernardi et al., 2016), especially deep learning methods (Goodfellow et al., 2016; Charniak, 2018) for generic *image captioning* (Fig. 1(a)) (Vinyals et al., 2017; Sharma et al., 2018; Hossain et al., 2019; Liu et al., 2019b; Donahue et al., 2015; Lu et al., 2017; Rennie et al., 2017), have recently led to increased research interest in DC (Jing et al., 2018; Li et al., 2019; Liu et al., 2019a; Kougia et al., 2019). However, despite its importance and recent



(a) Generic image captioning



FINDINGS: No change. No visible active cardiopulmonary disease. Both lungs remain clear and expanded. Heart and pulmonary XXXX are normal. No change in the large hiatus hernia.

SYSTEM TAGS: hiatus hernia / Hernia, Hiatus

HUMAN TAGS: Hernia, Hiatal/ large

(b) Diagnostic image captioning

Figure 1: (a) Caption produced by the *generic captioning* system of Vinyals et al. (2015). (b) Two images from an X-ray examination along with: the corresponding human-authored ‘FINDINGS’ section, from the IUXRay dataset (Section 2.1); silver labels (tags) automatically extracted from the human-authored findings (‘system tags’); and gold labels provided by physicians (‘human tags’). The diagnostic caption refers to ‘no change’ with respect to a previous examination, but there is no link to information and images of the previous examination in this particular dataset. The ‘XXXX’ is due to a (presumably automatic) de-identification process.

popularity, DC still suffers from shortcomings in methods, datasets, and evaluation measures. This article attempts to review all the published work on DC, outlining the major problems, and proposing future research directions.

DC methods usually employ the encoder-decoder architecture (Cho et al., 2014), heavily ignoring retrieval-based approaches. In a similar manner, Monshi et al. (2020) recently surveyed deep learning based DC methods, but only considered systems where the diagnostic text is generated from scratch, using a Recurrent Neural Network (RNN) decoder. However, recent studies show that retrieval-based approaches, where text from similar previous exams is reused, are very competitive despite their simplicity, often surpassing much more complex methods (Kougia et al., 2019). Even the simple nearest neighbor approach, where the diagnosis of the visually most similar study is retrieved and reused, has been reported to outperform all other approaches in clinical recall (Liu et al., 2019a). This could be due to the fact that retrieval can assist in capturing factual knowledge (Khandelwal et al., 2020) and negation; the latter is particularly important in clinical text (Kassner and Schütze, 2020). Or it could be due to the fact that medical reports seem to use a template-based language, where reports of the same findings are almost identical across different patients. We provide a segmentation of current DC systems by the kinds of methods they use, also reporting evaluation scores for each system with all available measures.

Datasets that have been used in previous DC work (Johnson et al., 2019; Demner-Fushman et al., 2015; Wang et al., 2018; Gale et al., 2018; Jing et al., 2018; Li et al., 2018) are not all publicly available. Out of the four publicly available ones, PEIR GROSS¹ and IMAGECLEF (de Herrera et al., 2018) suffer from severe shortcomings (Kougia et al., 2019), which are also discussed briefly below. Hence, in this work we focus mostly on studying and discussing the characteristics of the remaining two datasets, namely IU X-RAY (Demner-Fushman et al., 2015) and MIMIC-CXR (Johnson et al., 2019). Interestingly, previous research does not always use the same parts of the medical reports of these datasets (some use the ‘FINDINGS’ section only, others include the ‘IMPRESSION’), and most previous articles do not use a common training-development-test split of the data. For IU X-RAY, which is the most commonly used dataset, we use the split of Li et al. (2018), who recently used it to evaluate multiple DC systems. We also release (in supplementary material that will accompany the camera-ready) instructions on how to obtain and use this split.

Evaluation measures employed by previous DC research mainly assess lexical overlap between machine-generated and human-authored captions (Kougia et al., 2019), without directly assessing clinical correctness. This can lead to cases where a clinically wrong generated report can be scored higher than a clinically correct one (Zhang et al., 2019). Current methods for automatically measuring clinical correctness produce results of poor quality, because: (a)

¹PEIR Gross is a subcollection of the PEIR Digital Library (<https://peir.path.uab.edu/library/>).

they only consider the presence (or absence) of particular medical terms in the reports (Li et al., 2019), for example ‘pneumothorax’ would be considered a positive find in ‘no pneumothorax is observed’ (Zhang et al., 2019); or (b) they rely on the responses of rule-based automatic annotators (Liu et al., 2019a), for example to obtain the ‘system tags’ in Fig. 1(b), whose accuracy cannot be guaranteed; or (c) they use crowd workers (Li et al., 2018), who are not necessarily medical experts or trained in medical informatics.

Excluding an earlier version of this survey (Kougia et al., 2019), the only other DC survey we are aware of is that of Monshi et al. (2020).² As already pointed out, the latter considers only DC methods that generate diagnostic text from scratch using an RNN decoder, whereas we also consider retrieval-based methods, which are often very competitive in DC. We also scrutinize much more the datasets we consider. For instance, we study how often the diagnostic reports are very similar across patients, the class imbalance between reports with no findings vs. reports that report abnormalities, or to what extent relevant information is missing (e.g., reports referring to images that have been removed during anonymization, or sections that require access to unavailable previous examinations of the same patient). Furthermore, we provide a more extensive discussion of evaluation measures; for example, we also cover clinical correctness measures, apart from word overlap measures that Monshi et al. (2020) mostly focus on; and we demonstrate the shortcomings of current evaluation measures using concrete DC examples. A final difference from the survey of Monshi et al. (2020) is that we assume the reader is familiar with commonly used machine learning algorithms, including currently widely used deep learning (DL) models like Convolutional and Recurrent neural networks (CNNs, RNNs). This allows us to present and compare DC methods more succinctly. Readers who lack this background should consult introductory machine learning and DL textbooks first (Murphy, 2012; Goodfellow et al., 2016; Goldberg, 2017; Charniak, 2018).

2 Diagnostic Captioning Datasets

Datasets for DC comprise medical images and associated diagnostic reports. In previous work (Kougia et al., 2019), we reported that three publicly available datasets can be used for DC research, namely PEIR GROSS, ICLEFCAPTION (de Herrera et al., 2018), and IU X-RAY (Demner-Fushman et al., 2015). We concluded, however, that the first two datasets suffer from severe shortcomings. Most importantly, they contain photographs and captions from the figures of scientific articles, instead of real diagnostic medical images and reports; hence, they are inappropriate for realistic DC research. The third dataset, IU X-RAY, which contains X-ray images and medical reports, is appropriate if we ignore its small size. In our previous work (Kougia et al., 2019), we did not consider MIMIC-CXR (Johnson et al., 2019), a fourth and the largest to date publicly available DC dataset, which was released later, and we only partially explored IU X-RAY. In this work, we focus on the latter two quality datasets, MIMIC-CXR and IU X-RAY, referring readers interested in IMAGECLEF and PEIR GROSS to our previous work (Kougia et al., 2019). Datasets that do not comprise both medical images and diagnostic reports, or that are not publicly available, are not considered further in this study. Such datasets are BCIDR (Zhang et al., 2017b), consisting of 1,000 pathological bladder cancer images, each with five reports, which is not publicly available; Frontal Pelvic X-Rays (Gale et al., 2018), which comprises 50,363 images, each accompanied by a radiology report simplified to follow a standard template, but is not publicly available; and Chest X-Ray 14 (Wang et al., 2018), which is publicly available, but does not include any medical reports in its public version. Results on these datasets are included, however, in Table 4 for completeness.

Radiologists usually document their findings in titled sections, following standardized document structure templates. However, the sections found in the reports of IU X-RAY are not always the same as the sections found in the reports of MIMIC-CXR. ‘FINDINGS’, ‘COMPARISON’, ‘INDICATION’, ‘IMPRESSION’ are sections found in both datasets, among which ‘FINDINGS’ and ‘IMPRESSION’ are of primary interest (Johnson et al., 2019). The ‘FINDINGS’ section, which is usually the lengthier one (Fig. 2), describes the imaging characteristics of a body structure or function that can have a clinical impact. ‘COMPARISON’ contains previous information about the patient, often from preceding medical exams, but never the whole report or the medical images of the previous exams. This means that it is almost impossible even for a radiologist to generate this section without the previous referred exams. The same applies to the ‘INDICATION’ section, which conveys the medical reason for the patient to be subjected to the examination (e.g., symptoms). Hence, the ‘COMPARISON’ and ‘INDICATION’ sections cannot be generated by DC methods. Instead, they could be treated as given and they could, at least in principle, assist the process of generating the ‘FINDINGS’, although current DC methods attempt to generate the ‘FINDINGS’ directly from the images, without consulting the ‘COMPARISON’ and ‘INDICATION’ sections. All the aforementioned sections could in turn assist the process of generating the final section ‘IMPRESSION’, although again current DC methods try to generate the ‘IMPRESSION’ directly from the images. The ‘IMPRESSION’ usually summarizes the most important findings and interprets their

²The earlier version of this survey (Kougia et al., 2019) did not consider the recent MIMIC-CXR dataset, did not investigate as thoroughly the shortcomings of existing DC datasets and evaluation measures, and considered fewer (and less recent) DC methods, providing less information about them.

Table 1: Frequent ‘FINDINGS’ sections with no abnormality reported in IU X-RAY. Diagnostic findings with no abnormality reported (no abnormality tag assigned by experts) follow few templates. In IU X-RAY, for example, the top three most frequent ‘FINDINGS’ sections occur (exactly the same) in 148 reports (studies). By contrast, reports of abnormalities are less standardized.

Studies (#)	Reported Findings
51	The heart and lungs have XXXX XXXX in the interval. Both lungs are clear and expanded. Heart and mediastinum normal.
51	The heart is normal in size. The mediastinum is unremarkable. The lungs are clear.
46	Heart size normal. Lungs are clear. XXXX are normal. No pneumonia, effusions, edema, pneumothorax, adenopathy, nodules or masses.

Table 2: Number of reports per dataset (1st row) and of reports whose diagnostic text includes a section with Findings, Impression, Indication or Comparison (2nd to 5th row).

Number of Reports	IU X-RAY	MIMIC-CXR
# REPORTS	3851 (100%)	227,835 (100%)
# REPORTS WITH FINDINGS	3337 (87%)	149,758 (66%)
# REPORTS WITH IMPRESSION	3851 (99%)	187,793 (82%)
# REPORTS WITH INDICATION	3765 (98%)	155,607 (68%)
# REPORTS WITH COMPARISON	3252 (84%)	154,586 (68%)

clinical value, giving the referring physician a direction for the management of the disease or a final diagnosis. However, sometimes the ‘IMPRESSION’ (or ‘FINDINGS’) includes a conclusion that does not follow from the previous sections and the images of the current exam. For example, a conclusion may be the result of comparing the current exam with a previous one. Unfortunately, the dataset may omit the previous exam(s), as in IU X-RAY; or it may hide the dates and times of the exams of each patient, as in MIMIC-CXR, making it impossible to identify the previous exams.

The majority of the reports in both IU X-RAY and MIMIC-CXR concern cases where there is no disease or abnormality. In these cases, the diagnostic text is often very similar or identical across different exams (Table 1). The section that exists most often in both datasets is ‘IMPRESSION’ (Table 2). However, it is also the section with the shortest text on average (Fig. 2) and often comprises conclusions drawn from information not included in the datasets, as already noted. Some previous work used only the ‘IMPRESSION’ section as the target text to be generated (Shin et al., 2016a), but most previous work either uses the ‘FINDINGS’ as the target (Liu et al., 2019a; Li et al., 2018) or aims to generate the concatenation of the two sections (Shin et al., 2016a; Jing et al., 2018).

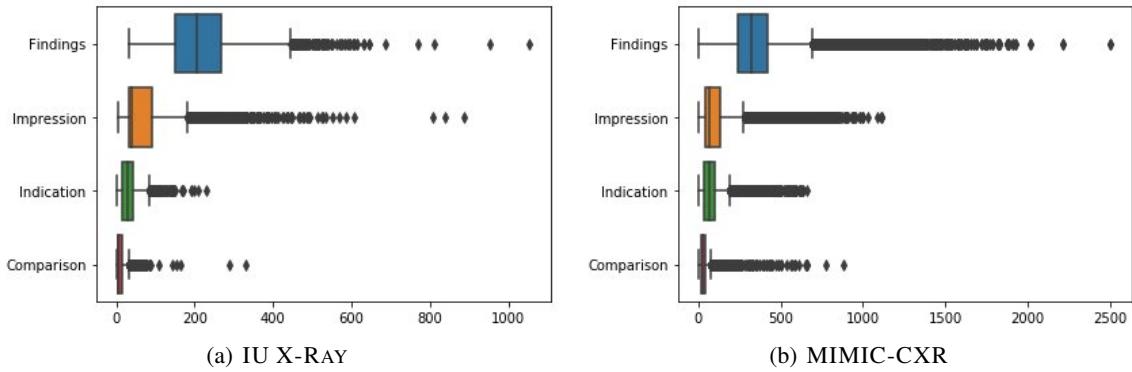


Figure 2: Boxplots of section lengths in characters, in (a) IU X-RAY and (b) MIMIC-CXR.

Another publicly available dataset is PadChest (Bustos et al., 2020), which comprises 160,868 chest X-Rays from 69,882 patients. However, its diagnostic texts (in Spanish) are not complete reports or complete sections of reports, but text snippets (not necessarily well formed sentences or paragraphs) that were extracted from the reports with regular expressions. Therefore, we exclude this dataset from our study, since it does not contain texts of the kinds the DC methods we consider aim to generate (entire reports or particular sections of reports).

2.1 IU X-RAY

Demner-Fushman et al. (2015) created a dataset of radiology examinations, comprising X-ray images and reports authored by radiologists. They publicly released an anonymized version of the dataset through the Open Access Biomedical Image Search Engine (OpenI).³ The dataset consists of 3,955 reports, one per patient, all in English, and 7,470 DICOM images.⁴ We found that 3,851 reports (97.4%) are linked to at least one image and are thus valid for our study. Among the 3,851 reports, 599 (15.6%) do not include a ‘COMPARISON’ section, 86 (2.2%) do not include the ‘INDICATION’ section, 31 (0.1%) do not include ‘IMPRESSION’, while 514 (13.3%) do not include ‘FINDINGS’. Among the 3,337 reports that comprise both a ‘FINDINGS’ section and at least one image, only 2,553 (76.5%) have a unique ‘FINDINGS’ section, i.e., the text of their ‘FINDINGS’ is not the same in any other report. The ‘FINDINGS’ section of the remaining 23.5% reports is exactly the same in two or more other reports, and in these cases the reports describe mainly normal findings. For example, the most frequent ‘FINDINGS’ text (in 51 reports) is: “*The heart is normal in size. The mediastinum is unremarkable. The lungs are clear.*” The 10 most frequent ‘FINDINGS’ sections, all describing normal findings, occur in 344 reports in total, which is 10.3% of all the reports and 43.9% of the non-unique ‘FINDINGS’. This gives an advantage to retrieval-based approaches, which have been reported to achieve surprisingly high performance overall (Kougia et al., 2019; Liu et al., 2019a; Boag et al., 2020).

Demner-Fushman et al. (2015) initially collected and de-identified 4k examinations from two hospitals, 2k from each one. They used software to de-identify the texts (Friedlin and McDonald, 2008), and the Clinical Trial Processor to de-identify the images.⁵ They reported excellent de-identification results (no sensitive information found by human annotators), but we found some cases where the software had damaged the diagnostic text. For example, “*Cardiomediastinal silhouette is XXXX*” is not informative and we cannot accurately infer the finding. Demner-Fushman et al. then discarded four exams that either did not comprise both a lateral and a posterioranterior chest image, or had diagnostic text that did not include clearly separated sections for the findings and the impression (or the diagnosis). They also discarded 41 exams containing information that could reveal the patient identities. Two of them revealed information such as address and dates, which the Health Insurance Portability and Accountability Act of 1996 (HIPAA) considers identifiers that may reveal personal information. The remaining 39 were exams with images that comprised teeth, partial jaw, jewelry, or partial skull. Although not explicitly flagged as sensitive by HIPAA, these images might identify the patients. These 39 exams had all of their images in this category and were completely removed from the dataset. However, 432 (10.9%) other exams had at least one image removed for the same reasons, but were apparently retained in the dataset (without the removed images). These 432 cases are problematic for training and evaluation purposes, because the gold (target) diagnostic text that the systems are required to generate may refer to an absent image.

The ‘IMPRESSION’ and ‘FINDINGS’ sections were used by Demner-Fushman et al. (2015) to manually associate each report with a number of tags, shown as ‘human tags’ in Fig. 1(b). Two human annotators were used, both trained in medical informatics. The tags were MeSH terms, supplemented with Radiology Lexicon terms (RadLex).⁶ Each annotation label (term) referred to a pathology, a foreign body or transplant, anatomy (human body parts), signs (imaging observations), or attributes (object or disease characteristics). The annotators were instructed not to assign labels for negated terms (e.g., ‘no signs of tuberculosis’) or inconclusive findings (e.g., introduced by ‘possibly’ or ‘maybe’, but not ‘probably’ or ‘likely’) of the ‘IMPRESSION’ and ‘FINDINGS’.⁷ In addition to the human tags, each report was associated with tags automatically extracted from the ‘IMPRESSION’ and ‘FINDINGS’ by Medical Text Indexer; the resulting tags are called ‘MTI encoding’ and they are shown as ‘system tags’ in Fig. 1(b). As shown in the example of Fig. 1(b), the system tags are single words or terms (e.g., ‘Hiatus’), while human tags follow a different pattern, which may combine anatomical site and type (e.g., ‘Hiatal/large’). Surprisingly, although the dataset comprises both human and system tags, only the latter have been used to evaluate (or train) DC systems so far; we discuss evaluation measures that use tags in Section 3.2 below.

2.2 MIMIC-CXR

MIMIC-CXR comprises 377,110 chest X-rays associated with 227,835 medical reports, from 64,588 patients of the Beth Israel Deaconess Medical Center examined between 2011 and 2016.⁸ The reports were de-identified to satisfy HIPAA requirements (Section 2.1). The images are chest radiographs, obtained from the hospital’s Picture Archiving

³<https://openi.nlm.nih.gov/>

⁴Each report is an XML file and images can be also downloaded in the PNG format. DICOM is a standard for medical images; see <https://www.dicomstandard.org/current/>.

⁵mirc.rsna.org/download/CTP-installer.jar

⁶Consult <https://www.nlm.nih.gov/mesh/meshhome.html> and <http://radlex.org/>.

⁷The annotation guidelines are publicly available (Demner-Fushman et al., 2015).

⁸When MIMIC-CXR was first introduced, these numbers were reported to be slightly different (Johnson et al., 2019). The numbers we provide refer to the v2.0.0. version that can be found at <https://mimic-cxr.mit.edu/>

and Communication System (PACS) in DICOM format. To remove all Protected Health Information (PHI), the images were processed to remove annotations imprinted in them (e.g., image orientation, anatomical position of the subject, timestamp of image capture). A custom algorithm was used for this purpose, based on image pre-processing and optical character recognition to detect text; all pixels within a bounding box containing this information were set to black. Two independent reviewers were employed upon de-identification to examine 6,900 radiographs for PHI, and they found no PHI.

The reports of MIMIC-CXR are written in English and their text is separated in sections, following document structure templates. Unlike IU X-RAY, where the boundaries of sections are made explicit by the XML markup, the section boundaries of MIMIC-CXR reports are not explicitly marked up. However, the section headings of MIMIC-CXR are written in upper case, followed by a column (e.g., ‘FINDINGS:’). Apart from the sections described in the discussion of IU X-RAY (Section 2.1), some reports of MIMIC-CXR include other sections, such as ‘HISTORY’, ‘EXAMINATION’, or ‘TECHNIQUE’, but not in a consistent manner, because the structure of the reports and section names were not enforced by the hospital’s user interface (Johnson et al., 2019).

3 Evaluation Measures for Diagnostic Captioning

Texts generated by DC systems have so far been assessed mostly via automatic evaluation measures originating from machine translation and text summarization (Papineni et al., 2002; Lin, 2004; Anderson et al., 2016; Vedantam et al., 2015), which, roughly speaking, count how many words or n -grams (phrases of n consecutive words) are shared between the generated text and reference gold texts (typically human-authored). Such measures have been reported to correlate well with human judgments of information content (e.g., the degree to which the most important information is preserved in summaries) when the goal is to rank systems and when there are multiple gold references per generated text. However, as can be seen in Table 3, measures of this kind do not necessarily capture clinical correctness. H_{ci}^a for example, which incorrectly reports ‘No pneumothorax’, receives higher scores than H_{cc}^a , which correctly reports pneumothorax. When the gold and the system-generated captions are automatically labelled for mentions of diseases or abnormalities (Xue et al., 2018; Huang et al., 2019; Liu et al., 2019a), standard classification evaluation measures, such as Accuracy, Precision and Recall, can be applied. Automated labelling, however, can be inaccurate leading to mistaken results. In the second (b) and third (c) cases of Table 3, for example, the automatically assigned tags are the same for the correct (H_{cc}) and incorrect (H_{ci}) diagnoses.

In an interesting evaluation approach, Li et al. (2019) employed crowd-sourcing for 100 randomly selected studies (examinations). For each study, the annotators were shown three reports, produced by a physician, a baseline DC system, and a more elaborate DC system, respectively. Each annotator had to consult the report of the physician and choose the best system-generated report, based on the criteria of clinical *correctness* of the reported abnormalities, *fluency*, and content *coverage* compared to the ground truth report. Although this approach is interesting in general, because it employs manual evaluation, the particular experiment of Li et al. (2019) raises doubts for three reasons. First, the medical background of the annotators was not reported and it may have been inadequate. Second, cases were excluded from the evaluation when no system-generated report was better than the other one, but these cases could be very frequent (e.g., if no system was good enough). Third, only 100 test studies were used and no statistical significance was reported. Consequently, we do not discuss the details and scores of this evaluation further.

3.1 Word Overlap Measures

The most common word overlap measures are BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005), which originate from machine translation and summarization, as already noted. The more recent CIDEr measure (Vedantam et al., 2015), which was designed for *generic* (not medical) image captioning (Kilickaya et al., 2016), has been used in only two DC works so far (Zhang et al., 2017b; Jing et al., 2018). SPICE (Anderson et al., 2016), also designed for generic captioning (Kilickaya et al., 2016), has not been used at all in DC so far. We note again that all these are word overlap measures, which do not always capture clinical correctness (Li et al., 2018), as already discussed. This was also discussed by Zhang et al. (2019), who used ROUGE to compare two medical statements, a clinically correct and a clinically incorrect one. Since the latter had more common words with the gold statement, it obtained a higher score. We included the example of Zhang et al. in Table 3, where we also used BLEU, METEOR, and ROUGE, along with Precision and Recall computed on CheXpert labels.

BLEU (Papineni et al., 2002) is the most common and the oldest among the word overlap measures that have been used in DC. It measures the word n-gram overlap between a generated and a ground truth caption. As this is a precision based measure, a brevity penalty is added to penalize short generated captions. BLEU-1 considers unigrams (i.e., single words), while BLEU-2, -3, -4 consider bigrams, trigrams and 4-grams respectively. The average of the four variants was used as the official measure in ICLEFCAPTION (Eickhoff et al., 2017; de Herrera et al., 2018). METEOR (Banerjee

Table 3: Using the BLEU (B1, B2, B3, B4), METEOR (Met), and ROUGE (Rou) automatic evaluation measures to score clinically correct (H_{cc}) and clinically incorrect (H_{ci}) hypothetical diagnoses that paraphrase three reference (gold, human) diagnoses (R^a, R^b, R^c). Percentage scores are reported. The first triplet (R^a, H_{cc}^a, H_{ci}^a) is from Zhang et al. (2019), while the rest were generated by a medical expert during our work. The last two columns show Precision (Pre) and Recall (Rec) scores of the ‘system’ tags of each H_{cc} and H_{ci} , which were extracted by the CheXpert labeller. All six automatic measures (B1, B2, B3, B4, Met, Rou) score the clinically incorrect diagnoses (H_{ci}) much higher than the clinically correct ones (H_{cc}).

TYPE	CLINICAL DIAGNOSIS	B1	B2	B3	B4	MET	ROU	TAGS	PRE	REC
R^a	Pneumothorax is seen. Bilateral pleural effusions continue.							Pneumothorax, Pleural Effusion		
H_{cc}^a	Pneumothorax is observed on radiograph. Bilateral pleural effusions continue to be seen.	58.3	46.1	34.9	26.2	46.1	66.3	Pneumothorax, Pleural Effusion	100.0	100.0
H_{ci}^a	No pneumothorax is observed. Bilateral pleural effusions continue.	75.0	65.5	52.3	41.1	55.2	81.0	Pleural Effusion	100.0	50.0
R^b	Stable cardiomegaly and mild bilateral interstitial opacities which represent mild pulmonary edema.							Cardiomegaly, Lung Opacity, Edema		
H_{cc}^b	Enlarged heart without notable variation and mild densities in both lungs, compatible with fluid accumulation.	13.3	9.8	0.0	0.0	5.8	15.1	Cardiomegaly	100.0	33.3
H_{ci}^b	Decreased cardiomegaly without considerable bilateral interstitial opacities, which exclude acute pulmonary edema.	58.3	46.1	34.9	26.2	29.0	58.3	Cardiomegaly	100.0	33.3
R^c	Mildly prominent perihilar opacities, due to bronchovascular crowding.							Lung Opacity		
H_{cc}^c	Bilateral interstitial densities in the hilar region, caused by increased bronchovascular markings.	8.3	0.0	0.0	0.0	10.4	10.4	NO FINDING	0	0
H_{ci}^c	There is no evidence of prominent perihilar opacities or bronchovascular crowding.	45.5	36.9	24.7	0.0	33.9	54.2	NO FINDING	0	0

and Lavie, 2005) extended BLEU-1 by employing the harmonic mean between Precision and Recall i.e., the F_β score, but biased towards Recall ($\beta > 1$). METEOR also employs Porter’s stemmer and WordNet (Fellbaum, 2012), the latter to take synonyms into account.⁹ The F_β score is then penalized up to 50% when no common n -grams exist between the machine-generated description and the reference human description. ROUGE-L (Lin et al., 2014) is the ratio of the length of the longest common n -gram shared by the machine-generated description and the reference human description, to the size of the reference description (ROUGE-L Recall); or to the generated description (ROUGE-L Precision); or it is a combination of the two (ROUGE-L F-measure). We note that many ROUGE variants exist (Graham, 2015), based on different n -gram lengths, stemming, stopword removal, etc., but ROUGE-L is the most commonly used version in DC so far. CIDEr (Vedantam et al., 2015) measures the cosine similarity between n -gram TF-IDF (Manning et al., 2008) representations of the two captions; words are also stemmed. Cosine similarities are calculated for unigrams to 4-grams and their average is returned as the final evaluation score. The intuition behind using TF-IDF is to reward terms that are frequent in a particular caption being evaluated, while penalizing terms that are common across captions (e.g., stopwords). However, DC datasets have constrained vocabularies, and a common disease name may thus be mistakenly penalized. More importantly, CIDEr scores exceeding 100% have been reported (Liu et al., 2019a), which contradicts the measure’s theoretic design. By using the official evaluation server implementation (CIDEr-D) (Chen et al., 2015), we found cases where scores exceeding 100% were indeed produced, which means that further investigation is required to check the correctness of the particular implementation of the measure and allow a fair comparison among systems. SPICE (Anderson et al., 2016) extracts sets of tuples from the two captions (human and machine-generated), containing objects, attributes, and/or relations; e.g., {(patient), (has, pain), (male, patient)}. Precision and recall are computed between the two sets of tuples, also taking WordNet synonyms into account, and the F_1 score is returned. The authors of SPICE report improved results over both METEOR and CIDEr, but it has been noted that results depend on the quality of syntactic parsing (Kilickaya et al., 2016). When experimenting with an implementation of this measure,¹⁰ we noticed that long texts were not parsed at times and thus were not evaluated properly.

3.2 Clinical Correctness Measures

The word overlap measures discussed above do not always capture clinical correctness, as already demonstrated. To overcome this problem, recent work has proposed new evaluation approaches based on classification evaluation measures, an approach we have already mentioned and we now discuss further. The clinical correctness of a generated caption is measured through a set of medical terms extracted from that caption (see Table 3). These terms are then compared to the ones from the gold caption, which may have been generated by humans (gold labels, as in IU X-RAY) or a system (silver labels), like the Medical Text Indexer (Mork et al., 2013) or the CheXpert labeller (Irvin et al., 2019). In Table 3 for example, CheXpert was used to annotate three reference diagnoses (R_a , R_b , R_c) along with their alternative correct and incorrect hypothetical diagnoses (H_{cc} , H_{ci}). In the topmost example, PLEURAL EFFUSION and PNEUMONOTHORAX have been correctly generated by CheXpert,¹¹ for the reference diagnosis and the correct hypothetical diagnosis H_{cc}^a . For the incorrect H_{ci}^a , only PLEURAL EFFUSION was generated, leading to a perfect Precision (number of correctly assigned tags to the total number of assigned tags) and a reasonable 50% Recall (number of correctly assigned tags to the number of gold tags). In the next example, CheXpert does not detect 2 out of 3 tags for H_{cc}^b , leading to a low 33.3% Recall. In the lowermost example, where the reference was labelled with LUNG OPACITY, no tags were detected by CheXpert for the correct hypothetical diagnosis H_{cc}^c . This leads to zero Recall and undefined Precision (though Precision is often taken to be also zero in such cases). Interestingly, however, the incorrect H_{ci}^c , which has the same (equally bad) Precision and Recall as H_{cc}^c , got high scores in many word overlap measures in Table 3, showing a weakness of such measures with respect to clinical accuracy assessment.

Xue et al. (2018) were the first to use an evaluation measure that considers medical tags extracted from system-generated and human-authored reports. The authors called the measure Keyword Accuracy, but it should not be confused with the conventional classification Accuracy, since it only measures Recall. The authors, who used the IU X-RAY dataset for their study, compiled a list of tags per examination and used it as the ground truth; the list consisted of the system-generated (MTI) tags and some of the human tags available in IU X-RAY. However, Xue et al. did not provide any further details (e.g., about the human tag selection criteria or how system and human tags were merged). Huang et al. (2019) followed the same approach, but they used only the MTI tags as their ground truth. In both of these works, however, where gold tags were compared with predicted tags, it is unclear how the predicted tags were extracted from the system-generated reports. Liu et al. (2019a) used the CheXpert medical abnormality mention detection system (Irvin et al., 2019), which generates one out of 4 labels (presence, absence, negative, not sure) for each one of 14 thoracic diseases. For any given report, any disease for which the assigned label was ‘presence’ was considered to be mentioned in that report. When the assigned label for a disease was ‘not sure’, then the authors considered that

⁹<https://tartarus.org/martin/PorterStemmer/>

¹⁰<https://goo.gl/b011Bz>

¹¹We followed the work of Liu et al. (2019a), who flipped a coin when CheXpert returned the ‘unsure’ label.

the disease was mentioned in the report with 0.5 probability. This process was applied to both system-generated and gold reports and then micro- and macro-averaged Precision and Recall were computed (along with macro-averaged Accuracy). A disadvantage of this approach is that it uses only 14 diseases, which is a very small number compared to the hundreds of abnormality tags of other works and the much wider variety of medical conditions physicians need to consider in practice. However, the work of Liu et al. (2019a) can be used to highlight the limitations of Accuracy compared to Precision and Recall, when used to assess DC systems. A majority classifier (a system that always reported no findings, the majority prediction) obtained a higher Accuracy than more elaborate methods in the experiments of Liu et al. More generally, it is well known in Machine Learning that a large class imbalance may lead to misleadingly high Accuracy, which is why Precision, Recall, and F1 are used instead in such cases. In Table 4, we computed and report the harmonic mean (F1) of Precision and Recall using the results of Liu et al. (2019a). We note that Receiver Operating Characteristics (ROC) curves, which plot the true positive rate against the false positive rate for different classification probability cut-offs, can also be used to get a better view of the performance of systems and baselines (Swets, 1988). Precision-Recall (PR) curves can be used in a similar manner. In both cases, the Area Under each Curve (AUC) can serve as a single evaluation score that aggregates results over different cut-offs. However, previous DC work does not provide enough information to reconstruct ROC or PR curves, and does not report AUC scores.

4 Diagnostic Captioning Methods

We now discuss the main types of DC methods, including their relation to generic image captioning. We also briefly cover early approaches that did not process medical images directly, but were fed with findings manually extracted from medical images, or that did not generate text, but were intended to help in the manual preparation of diagnostic reports.

Early approaches

Varges et al. (2012) followed an ontology-based natural language generation approach to assist medical professionals turn cardiological findings (from diagnostic images) into readable and informative textual descriptions. The input to the text generator, however, was not directly a medical image, but triplets like *<right atrium, size, normal>*. From a different perspective, Schlegl et al. (2015) used medical images and their diagnostic reports as input to a Convolutional Neural Network (CNN) to classify voxels (3D pixels) as intraretinal cystoid fluid, subretinal fluid, or normal retinal tissue, with the help of concepts automatically extracted from the text of the corresponding report; in this case, the report was part of the system’s input, not the system’s output. Kisilev et al. (2015a,b) performed semi-automatic lesion detection and contour extraction from medical images. Structured Support Vector Machines (Tsochantaridis et al., 2004) were then used to generate semantic tags, originating from a radiology lexicon, for each lesion. In later work, Kisilev et al. (2016) used a CNN to detect Regions of Interest (ROIs) in the images, then fully connected layers to assign predefined features describing abnormalities to each ROI. The assigned features were finally filled into sentence templates to generate captions. We discuss template-based text generation below.

Generic Image Captioning vs. Diagnostic Captioning

Deep learning approaches are currently dominant in both generic and diagnostic image captioning.¹² Hossain et al. (2019) compiled a taxonomy of aspects of *generic* image captioning methods that are based on deep learning. Fig. 3 depicts that taxonomy. We highlight aspects that have *not* also been used in *diagnostic* captioning work. The fact that most of the aspects have also been used in DC indicates that generic image captioning methods are also applicable (or at least have been considered) in DC. The best generic image captioning methods, however, are not necessarily the best ones for DC, mostly because of two factors. First, DC methods do not aim to simply describe what is present in an image, unlike generic image captioning methods. DC aims to report clinically important information that is relevant for diagnostic purposes. Simply reporting, for example, which organs are shown in a medical image is undesirable, if there is nothing clinically important to be reported about them. Second, as we have already discussed in previous sections, diagnostic reports are often very similar across examinations of different patients. This allows retrieval-based approaches to perform surprisingly well, often challenging encoder-decoder approaches that are currently the state of the art in generic image captioning. We present both approaches and other alternatives below.

Regarding the type of input ('feature mapping' box of Fig. 3) that is used to generate diagnostic captions, both images (Li et al., 2018; Wang et al., 2018; Gale et al., 2018; Xue et al., 2018; Liu et al., 2019a; Huang et al., 2019) and images combined with text (Shin et al., 2016b; Jing et al., 2018; Yin et al., 2019; Yuan et al., 2019) have been explored. Both supervised and reinforcement learning have been employed in DC (Jing et al., 2018; Li et al., 2018; Liu et al., 2019a); the latter falls in the 'other' category of learning type in the taxonomy of Fig. 3. A caption in current DC work

¹²For a survey of earlier work on generic image-to-text generation, consult Bernardi et al. (2016).

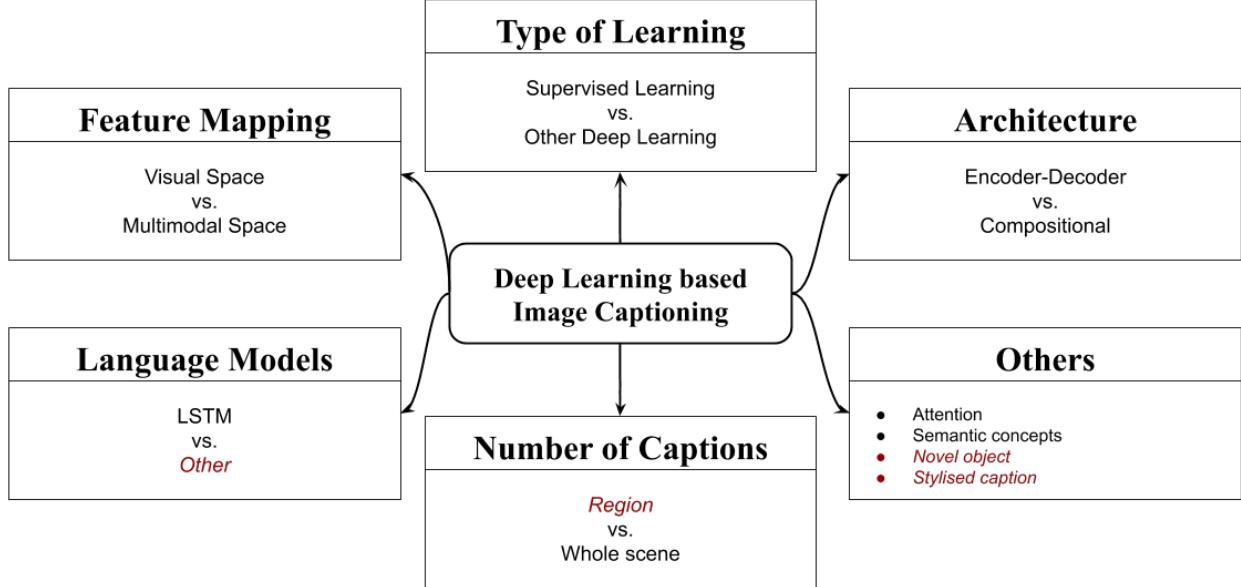


Figure 3: Aspects of deep learning generic image captioning methods, using the terminology of Hossain et al. (2019). Aspects that have *not* been used in DC are shown in red and italics.

typically refers to an entire medical image or even to a set of medical images, for example multiple X-rays from a single examination (Li et al., 2018; Yuan et al., 2019), not to a particular region of the image(s), unlike some generic image captioning work (Hossain et al., 2019). The most common system architecture in DC is the encoder-decoder approach. However, other approaches (dubbed ‘compositional’ by Hossain et al. (2019)) have also been tried in DC, for example the knowledge graph approach by Li et al. (2019). Usually, LSTMs (viewed as language models in the taxonomy of Fig. 3) are employed to generate text in DC, but we observe that Transformer-based models (Devlin et al., 2018; Radford et al., 2019) can also be employed, which is a direction with potential that has just begun being explored (Chen et al., 2020). Regarding the ‘other’ box of Fig. 3, where miscellaneous other aspects are listed, concepts (Jing et al., 2018) and attention-based methods are common in DC (Wang et al., 2018; Jing et al., 2018; Gale et al., 2018; Li et al., 2018; Xue et al., 2018; Liu et al., 2019a; Huang et al., 2019; Yin et al., 2019; Yuan et al., 2019). Gaze mechanisms, which have not been explored in generic image captioning (Hossain et al., 2019), could also be useful in DC. For example, attention mechanisms might aim to mimic how a physician focuses in turn on different parts of a medical image. By contrast, the emotional aspects that have been studied for generic image captioning (Nezami et al., 2020), are not related to DC. ‘Novel objects’ in generic image captioning are objects present only in the test dataset (Agrawal et al., 2019). Novel object captioning has not been investigated in DC yet. However, it would certainly be interesting to assess system-generated text for patients with rare (or new) conditions. ‘Stylised captioning’, which aims to generate stylised and attractive descriptions of generic images, is of little importance in DC, where informativeness and clinical accuracy matter most.

Diagnostic Captioning Architectures

Most often, the encoder-decoder architecture (Cho et al., 2014) is employed in DC (Bai and An, 2018; Liu et al., 2019b), with or without visual attention and reinforcement learning. Systems that adopt this architecture first encode the medical images as dense vectors, typically using CNN-based image encoders. They then generate the diagnostic text from the image encoding, typically using Recurrent Neural Network (RNN) decoders. However, retrieval-based methods have also been proposed for DC (Li et al., 2018), and even their simplest forms (e.g., reusing the report of the visually nearest training instance) have been found to outperform all other systems in clinical (tag-based) Recall (Liu et al., 2019a). As shown in Table 4, more elaborate retrieval-based systems can outperform state of the art encoder-decoders in DC. More specifically, the retrieval-based system of Li et al. (2019) achieved overall better results than their earlier hybrid encoder-decoder and retrieval-based approach (Li et al., 2018). The reader is warned that not all of the results of Table 4 are directly comparable, since some of them are obtained from different datasets, or different training/development/test splits. However, the very high manual evaluation score of the retrieval-based method of Li et al. (2019) is an indication that the encoder-decoder approach may be worse than retrieval-based approaches to DC, and that the latter should be explored more in DC. We also note that retrieval has been recently found to improve language models (Kassner and

Table 4: Evaluation scores of DC methods, using BLEU-1/-2/-3/-4 (B1, B2, B3, B4), METEOR (M), ROUGE-L (R), CIDEr (C), manual evaluation (ME), clinical F1 (CF1). All scores are percentages. The first column shows the name of the system or the authors who introduced it. The FROM column shows the article the results were obtained from. For MIMIC-CXR, we computed CF1 from the micro-averaged Precision and Recall reported by Liu et al. (2019a). The TYPE column shows the type of method: Retrieval-based (Rb), Encoder-Decoder (ED), or Decoder only (D). The columns VA, HD, RL correspond to visual attention, hierarchical decoding, reinforcement learning, respectively; these are some of the most frequent aspects of ED methods. Datasets marked with \dagger are not publicly available. Scores for the same dataset are strictly comparable only when obtained from the same article (FROM column), because of different training/development/test splits that may have been used otherwise. The first two systems are baselines. For all other systems, when multiple versions exist, the performance of the most extensive version is reported, as opposed to simpler versions used, for example, in ablation studies.

SYS./AUT.	TYPE	VA	HD	RL	FROM	DATASET	B1	B2	B3	B4	M	R	C	ME	CF1		
1-NN	RB				Liu et al. (2019a)	IU X-RAY	23.2	11.6	5.1	1.8	20.1	72.8			39.1		
					Liu et al. (2019a)	MIMIC-CXR	30.5	17.1	9.8	5.7	24.4	75.5					
BlindRNN	D				Liu et al. (2019a)	IU X-RAY	23.3	13.0	8.7	6.1	29.1	74.7					
					Liu et al. (2019a)	MIMIC-CXR	26.9	17.2	11.3	7.4	27.2	71.6					
Xue et al. 2018	ED	✓	✓		Xue et al. (2018)	IU X-RAY	46.4	35.8	27.0	19.5	27.4	36.6					
Huang et al. 2019	ED	✓	✓		Huang et al. (2019)	IU X-RAY	47.6	34.0	23.8	16.9	34.7	29.7					
Shin et al. 2016b	ED				Shin et al. (2016b)	IU X-RAY	78.5	14.4	4.7	0.0							
Wang et al. 2018	ED	✓			Wang et al. (2018)	CHEST X-RAY 14 \dagger	28.6	15.9	10.3	7.3	10.7	22.6				40.6	
					Liu et al. (2019a)	IU X-RAY	33.0	19.4	12.4	8.1	31.1	133.4					
					Liu et al. (2019a)	MIMIC-CXR	33.2	21.2	14.2	9.5	29.6	100.4					
Zhang et al. 2017b	ED	✓			Zhang et al. (2017b)	BCIDR \dagger	91.2	82.9	75.0	67.7	39.6	70.1	2.04				
Gale et al. 2018	ED	✓			Gale et al. (2018)	PELVICX \dagger	91.9	83.8	76.1	67.7							
					Gale et al. (2018)	PELVICX TT \dagger	65.0	37.9	24.2	15.9							
Liu et al. 2019a	ED	✓	✓	✓	Liu et al. (2019a)	IU X-RAY	35.9	23.7	16.4	11.3	35.4	142.4				33.8	
					Liu et al. (2019a)	MIMIC-CXR	31.3	20.6	14.6	10.3	30.6	104.6					
Jing et al. 2018	ED	✓	✓		Li et al. (2019)	IU X-RAY	45.5	28.8	20.5	15.4	36.9	27.7	24.1				
					Jing et al. (2018)	PEIR GROSS	30.0	21.8	16.5	11.3	14.9	27.9	32.9				
Li et al. 2018	ED/RB	✓	✓	✓	Li et al. (2019)	IU X-RAY	43.8	29.8	20.8	15.1	32.2	34.3	48.0				
					Li et al. (2019)	CX-CHR \dagger	67.3	58.7	53.0	48.6	61.2	289.5					
Li et al. 2019	RB		✓		Li et al. (2019)	IU X-RAY	48.2	32.5	22.6	16.2	33.9	28.0	57.4				
					Li et al. (2019)	CX-CHR \dagger	67.3	58.8	53.2	47.3	61.8	285.0	67.8				
Vinyals et al. 2015	ED				Liu et al. (2019a)	IU X-RAY	26.5	15.7	10.5	07.3	30.6	92.6				33.1	
					Liu et al. (2019a)	MIMIC-CXR	30.7	20.1	13.7	09.3	30.0	88.6					
Xu et al. 2015	ED	✓			Liu et al. (2019a)	IU X-RAY	32.8	19.5	12.3	08.0	31.3	127.6					
					Liu et al. (2019a)	MIMIC-CXR	31.8	20.5	13.7	9.3	28.8	96.7	39.6				
Rennie et al. 2017	ED	✓		✓	Li et al. (2019)	IU X-RAY	22.4	12.9	08.9	06.8	30.7	29.7					
Lu et al. 2017	ED	✓			Li et al. (2019)	IU X-RAY	22.0	12.7	08.9	06.9	30.8	29.6					
Yuan et al. 2019	ED	✓	✓		Yuan et al. (2019)	IU X-RAY	52.9	37.2	31.5	25.5	34.3	45.3					
Yin et al. 2019	ED	✓	✓		Yin et al. (2019)	IU X-RAY	44.5	29.2	20.1	15.4	17.5	34.4	34.2				

Schütze, 2020; Khandelwal et al., 2020). The benefits may be greater when modeling the language of diagnostic reports, where large parts of text are often very similar or exactly the same across different patients, as already discussed.

Having provided a brief overview of the most common DC approaches, we now discuss them in more detail, starting from the encoder-decoder architecture.

Encoder-Decoder (ED) The encoder-decoder deep learning architecture was originally introduced for machine translation (Cho et al., 2014), but was then also adopted in generic image captioning (Hossain et al., 2019; Bai and An, 2018; Liu et al., 2019b). In machine translation or other text-to-text generation tasks (e.g., summarization) an encoder network, often an RNN such as an LSTM (Hochreiter and Schmidhuber, 1997), reads the input text and converts it to a single vector or a sequence of vectors. A decoder network, often another RNN, then produces the target text, in the simplest case word by word, using as its input the encoding of the input text. An attention mechanism (Xu et al., 2015) allows the decoder to focus on particular vectors of the input text encoding, if the latter is an entire sequence of vectors. In generic image captioning, the encoder is typically a CNN (LeCun et al., 2015), which converts the image into a single or multiple vectors (e.g., corresponding to patches of the image). The decoder again produces the target text (caption), using the image encoding as its input. An attention mechanism may again allow the decoder to focus on particular vectors of the image encoding when generating each word; we call mechanisms of this kind ‘visual attention’ and we discuss them separately below. An example of an encoder-decoder model without visual attention for generic

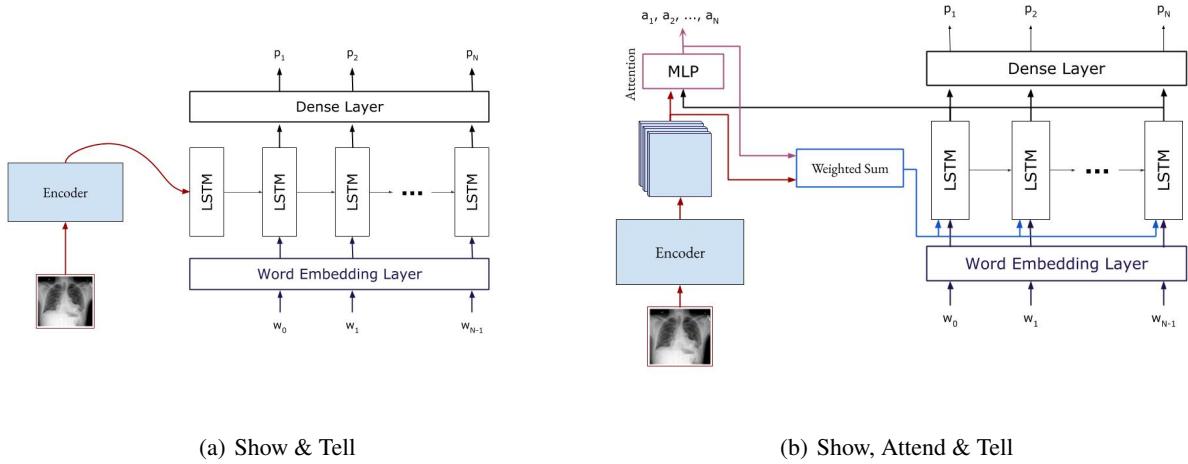


Figure 4: Left: the Show & Tell (S&T) model by Vinyals et al. (2015). Right: the Show, Attend & Tell (SA&T) model by Xu et al. (2015). S&T uses the image encoding of the CNN to initialise the LSTM decoder. SA&T also comprises a visual attention mechanism.

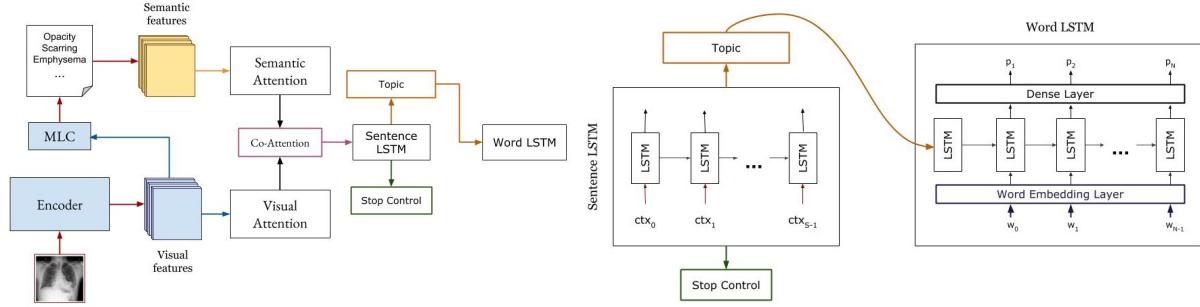
image captioning is the model of Donahue et al. (2015). This model comprises a CNN to encode the image and an LSTM to decode to text. The CNN was CaffeNet (Jia et al., 2014) or the better performing VGG (Simonyan and Zisserman, 2014). The decoder was an LSTM, which used the representation of the previous generated word along with the image encoding to generate the next word at each timestep. The authors also experimented with a stacked, two-layer LSTM decoder. Another example is the well-known Show & Tell (S&T) system of Vinyals et al. (2015), which was also introduced for generic image captioning; see Fig. 4(a). It employs the Inceptionv3 CNN (Szegedy et al., 2016) to encode the image and uses the image encoding to initialise the LSTM decoder.

ED + Visual Attention (VA) We place in this category encoder-decoders that also employ visual attention mechanisms (VA), as in Fig. 4(b) (Xu et al., 2015). Such mechanisms can also be used to highlight on the image the findings described in the report to make the diagnosis more easily interpretable (Zhang et al., 2017b; Jing et al., 2018; Wang et al., 2018; Yin et al., 2019; Yuan et al., 2019). Zhang et al. (2017b) were the first to employ visual attention in DC with the MDNet model.¹³ They used the BCIDR dataset (not publicly available, Section 2), which contains pathological bladder cancer images and diagnostic reports, aiming to generate paragraphs conveying findings. MDNet used a form of ResNet (He et al., 2016) to encode images. The image encoding acts as the initial hidden state of an LSTM decoder, which also uses visual attention. The decoder was cloned to generate multiple sentences. However, in most evaluation measures the model performed only slightly better than the generic image captioning model of Karpathy and Fei-Fei (2015) applied to DC.

The system of Lu et al. (2017), which was designed for generic image captioning and uses visual attention too, was also applied to DC (Li et al., 2018). Its CNN encoder is a ResNet (He et al., 2016), and its decoder is an LSTM. At each timestep, the spatial image encodings (one per image region) and the LSTM hidden state are used as input to a Multi-Layer Perceptron (MLP) with a single hidden layer and a softmax output activation function, acting as a visual attention mechanism (Xu et al., 2015). This mechanism generates one weight per image region, and the weights are used to form an overall weighted image representation. This image representation is then used along with the hidden state of the LSTM decoder to predict the next word. The authors also extended the decoder with a binary gate, which allows deactivating the visual attention when visual information is redundant (e.g., generating stopwords may require no visual attention).

Jing et al. (2018) created an encoder-decoder model with visual attention especially for DC, illustrated in Fig. 5(a). They used VGG-19 (Simonyan and Zisserman, 2014) to encode each image and extract equally sized patches. Each patch encoding is treated as a ‘visual’ feature vector. An MLP, called MLC in their article and in Fig. 5(a), is then fed with the visual feature vectors and predicts terms from a pre-determined term vocabulary. The word embeddings (dense vector representations) of the predicted terms of each image are treated as ‘semantic’ feature vectors representing the image. The decoder, which produces the text, is a hierarchical RNN, consisting of a sentence-level LSTM and a

¹³Zhang et al. introduced TandemNet (Zhang et al., 2017a) earlier, which also used visual attention, but for medical image classification. TandemNet could also perform captioning, but the authors focused on this task in MDNet.



(a) Image encoder, visual and semantic features, attention mechanisms, and hierarchical decoder.

(b) Hierarchical text decoder, consisting of a sentence-level LSTM and a word-level LSTM.

Figure 5: Jing et al. (2018) proposed a DC model that first encodes the image, then extracts ‘visual’ features from the image encoding and ‘semantic’ features from terms predicted from the image encoding. Attention mechanisms are used to produce an overall image representation from the visual and semantic features, which is different at each timestep of the sentence LSTM decoder. At each timestep, the sentence decoder selects the topic of the corresponding sentence. A word LSTM decoder then generates the words of the sentence.

word-level LSTM. The sentence-level LSTM produces a sequence of sentence embeddings (vectors), each intuitively specifying the information to be expressed by a sentence of the image description (acting as a sentence topic). For each sentence embedding, the word-level LSTM then produces the words of the corresponding sentence, word by word. More precisely, at each one of its timesteps, the sentence-level LSTM examines both the visual and the semantic feature vectors of the image. An attention mechanism (an MLP fed with the current state of the sentence-level LSTM and each one of the visual feature vectors of the image) assigns attention scores to the visual feature vectors, and the weighted sum of the visual feature vectors (weighted by their attention scores) becomes a visual ‘context’ vector, intuitively specifying which visual features to express by the next sentence. Another attention mechanism (another MLP) assigns attention scores to the semantic feature vectors (representing the image terms), and the weighted sum of the semantic feature vectors (weighted by attention) becomes the semantic context vector, specifying which terms of the image to express by the next sentence. At each timestep, the sentence-level LSTM considers the visual and semantic context vectors, produces a sentence embedding (topic), and updates its state, until a stop control instructs it to stop. Given a sentence embedding, the word-level LSTM produces the words of the corresponding sentence, until a special ‘stop’ token is generated. Jing et al. showed that their model outperforms generic image captioning models with visual attention (Vinyals et al., 2015; Donahue et al., 2015; Xu et al., 2015; You et al., 2016) in DC. Wang et al. (2018) adopted an approach to DC similar to that of Jing et al. (2018), using a ResNet-based image encoder and an LSTM decoder, but their LSTM is flat, as opposed to the hierarchical LSTM of Jing et al. (2018). Wang et al. (2018) also extract additional image features from the states of the LSTM.

Yin et al. (2019) created an encoder-decoder DC model similar to that of Jing et al. (2018). Again, a hierarchical LSTM attends over image features and representations of abnormality labels predicted from the image encoding. The image encoder is DenseNet (Huang et al., 2017), but Yin et al. remove its last global pooling layer (arguing it could lose important spatial information) and the last fully connected layer (which serves as a classifier). Instead they add a convolutional layer that operates on the image region representations and outputs a probability distribution for each particular label over the image, intuitively a heatmap for each label. A global max pooling over each label’s heatmap then produces a single probability per label for the entire image. The DenseNet encoder was pre-trained on ImageNet and then fine-tuned on IU X-ray (Section 2). The hierarchical LSTM and the attention mechanisms are very similar to those of Jing et al. (2018) discussed above. Yin et al. also add a ‘topic matching’ loss that roughly speaking penalizes topic representations (sentence embeddings) produced by the sentence-level LSTM decoder when they deviate from the representations of the corresponding ground-truth sentences.

Another encoder-decoder model with visual attention for DC was proposed by Yuan et al. (2019). It uses the ResNet-152 image encoder (He et al., 2016), which Yuan et al. pre-trained on the medical image dataset CheXpert (Irvin et al., 2019) to perform multi-label classification with 14 labels (12 disease labels, “Support Devices” and “No Finding”).¹⁴ The image encoder was fine-tuned by training it for the same task on IU X-ray (Section 2) and then used with a new dense layer added on top for classification with 69 gold labels (medical concepts) extracted from the ground truth

¹⁴CheXpert is both the name of a dataset (containing 224,316 chest radiographs) and the name of a tool (‘CheXpert labeler’) that was used to annotate the dataset. The latter tool was also used in Table 3.

reports by SemRep.¹⁵ Yuan et al. show experimentally that an image encoder pre-trained on a large dataset of medical images (224,316 chest X-rays) has better performance than encoders pre-trained on ImageNet. The decoder is again a hierarchical LSTM, in which the sentence LSTM attends over the image and the word LSTM over the medical concepts produced by the encoder. Yuan et al. report state-of-the-art results, outperforming Li et al. (2019) and Jing et al. (2018) among others. Yuan et al. also allow their model to be fed with multiple images to generate a single report from. This is important, because many imaging examinations comprise multiple images, for example a frontal and a lateral projection image. Systems trained to generate a report from a single medical image at a time cannot handle well such cases. Similar provisions are made in the system of Li et al. (2018), which also uses Reinforcement Learning, discussed below.

ED + Reinforcement Learning (RL) These methods use the encoder-decoder architecture, but also employ Reinforcement Learning (RL) (Sutton and Barto, 2018). For example, Rennie et al. (2017) employed the REINFORCE algorithm (Williams, 1992) with a reward based on CIDEr (Section 3.1), but in the context of generic image captioning. An advantage of RL is that non-differentiable evaluation measures can be used directly during training, so that systems are not optimizing loss functions like cross-entropy during training while being assessed with measures such as BLEU, ROUGE, or clinical F1 (Section 3) at test time. For readers not familiar with these issues, we note that when training with backpropagation the loss function must be differentiable, which is not the case for most current DC evaluation measures. By contrast with RL the reward does not need to be differentiable. It can also be given at the end of a sequence of system decisions, in cases where a loss is not available for each individual decision. In the DC system of Li et al. (2018), for example, RL is used to decide if a sentence will be generated from scratch, or if it will be retrieved from a database with frequently occurring sentences. The image encoding, produced by a DenseNet-121 (Iandola et al., 2014) or a VGG-19 (Simonyan and Zisserman, 2014) CNN, is fed to a hierarchical RNN decoder similar to that of Jing et al. (2018), illustrated in Fig. 5(b). However, for each sentence embedding (topic) produced by the sentence-level RNN, an agent trained using RL (again using REINFORCE and CIDEr) decides if the sentence will be generated using the word-level RNN or if it will be generated by using a sentence retrieved from a database of frequent sentences. Li et al. (2018) applied their system to IU X-RAY (Section 2), but their experimental results were close to those of a baseline. In more recent work, Liu et al. (2019a) used DenseNet-121 (Iandola et al., 2014) for image encoding and a hierarchical LSTMs decoder. Similarly to Li et al. (2018) and Rennie et al. (2017), REINFORCE with a CIDEr-based reward was employed. However, this time RL was used to optimise readability. Liu et al. also included a reward based on comparing labels, like the ones of Fig. 1(b), extracted by CheXpert (Irvin et al., 2019) from the system-generated text and the human-authored report, in order to optimize clinical accuracy.

ED + Language Templates (LT) Template-based generation has a long history in natural language generation (Reiter and Dale, 2000; Gatt and Krahmer, 2018), where templates of many different forms have been used, ranging from surface-form sentence templates, to sentence templates at the level of syntax trees, to document structure templates (Van Deemter et al., 2005). In the context of DC, language templates (LT) have recently been combined with encoder-decoder approaches, attempting to provide more satisfactory diagnostic reports. Gale et al. (2018) focused on classifying hip fractures in frontal pelvic X-rays, and argued that generating reports for such narrow medical tasks can be simplified to using only two sentence templates; one for positive cases, including five placeholders (slots) to be filled in by descriptive terms, and a fixed negative template with no slots. They used DenseNet (Huang et al., 2017) to encode the image, and (presumably) classify it as a positive or negative case, and a two-layer LSTM with attention over the image encoding to fill in the slots of the positive template. Their scores are very high (Table 4), but this is expected due to the extremely simplified and standardized ground truth reports. For example, the vocabulary of the latter contains only 30 words, including special tokens. The DC systems of Li et al. (2018) and Liu et al. (2019a) also use sentence templates, but these are rather complete sentences, with no empty slots to fill.

Retrieval-based approaches to DC can be as simple as reusing the diagnostic text of the visually nearest (in terms of image encoding similarity) medical exam of the training set (Liu et al., 2019a). Even this 1-nearest neighbor approach achieves surprisingly good results; see the second row of Table 4. Then, it should be no surprise that the more advanced retrieval-based DC approach of Li et al. (2019) outperforms ED (Vinyals et al., 2015; Donahue et al., 2015), ED+VA (Jing et al., 2018), and ED+RL (Li et al., 2018) methods in Table 4 (we employ comparable results from the Table). We also note that methods that retrieve sentences (Li et al., 2018, 2019), discussed above, can also be seen as belonging in the category of retrieval-based systems. Retrieval-based systems were also the top performing submissions of the ImageCLEF Caption Prediction subtask, a task that ran for two consecutive years (Eickhoff et al., 2017; de Herrera et al., 2018).¹⁶ The top participating systems of the competition in both years relied on (or included) image retrieval (Liang et al., 2017; Zhang et al., 2018). Zhang et al. (2018), who obtained the best results in 2018, used the Lucene Image Retrieval system (LIRE) to retrieve similar images from the training set, then simply concatenated the captions

¹⁵<https://semrep.nlm.nih.gov/>

¹⁶Later ImageCLEF Caption tasks (Pelka et al., 2019, 2020) only required systems to assign labels to medical images, without requiring diagnostic text to be generated.

of the top three retrieved images to obtain the new caption.¹⁷ Liang et al. (2017), who had the best results in 2017, combined an ED approach with image-based retrieval. They reused a pre-trained VGG encoder and an LSTM decoder, similarly to those of Karpathy and Fei-Fei (2015). They trained three such models on different caption lengths and used an SVM classifier to choose the most suitable decoder for the given image. They also used a 1-nearest neighbor method to retrieve the caption of the most similar training image and concatenated it with the generated caption.

Baselines are included in the first two lines of Table 4. BlindRNN is an RNN that simply generates word sequences, having been trained as a language model on medical captions, without considering the image(s); a single-layer LSTM was used in the BlindRNN of Table 4. The 1-NN baseline retrieves the diagnostic text of the visually most similar image from the training set (Liu et al., 2019a). These simplistic baselines were intended to be easy to beat, but as can be seen in Table 4, the scores of 1-NN are very high, and they outperform some much more elaborate approaches, such as the system of Liu et al. (2019a) in clinical recall.

5 Conclusions and Directions for Future Research

We have provided an extensive overview of diagnostic captioning (DC) methods, publicly available datasets, and evaluation measures.

In terms of methods, most current DC work uses encoder-decoder deep learning approaches, largely because of their success in generic (non-medical) image captioning. We have pointed out, however, that DC aims to report only information that helps in a medical diagnosis. Prominent objects shown (e.g., body organs) do not need to be mentioned, if there is nothing clinically important to be reported about them, unlike generic image captioning where salient objects (and actions taking place) typically have to be reported. Another major difference from generic image captioning is that medical images vary much less and, consequently, the corresponding diagnostic text is often very similar or even identical across different patients. These two factors allow retrieval-based methods, which reuse diagnostic text from training examples with similar images, to perform surprisingly well in DC. Frequent sentences or sentence templates can also be used, instead of generating them.

For evaluation purposes, DC work has so far relied mostly on word overlap measures, originating from machine translation and summarization, which often fail to capture clinical correctness, as we have also demonstrated using artificial examples. Measures that compare tags (also viewed as labels or classes, corresponding to medical terms or concepts) that are manually or, more often, automatically extracted from system-generated and human-authored diagnostic reports have also been employed, as a means to better capture clinical correctness. They may also fail, however, when the tools that automatically extract the tags are inaccurate, when human annotation guidelines are unclear on exactly which tags should be assigned or not, and when tags cannot fully capture the information to be included in the diagnostic text. Manual evaluation is rare in DC, presumably because of the difficulty and cost of employing evaluators with sufficient medical expertise.

In terms of datasets, we focused on the only two publicly available datasets that are truly representative of the task (IU X-RAY, MIMIC-CXR), having first discussed severe shortcomings of the other publicly available datasets (e.g., they may not contain medical images from real examinations). We also collected and reported evaluation results from previous published work for all the DC datasets, methods, and evaluation measures we considered. Although these results are often not directly comparable, because of different datasets or splits used, they provide an overall indication of how well different types of DC methods perform. The results we collected may also help other researchers produce results that will be more directly comparable to previously reported ones.

Our main findings also guide our proposals for future work on DC, listed below.

Hybrid explainable DC-specific methods. We believe that hybrid methods, which will combine encoder-decoder approaches that generate diagnostic text from scratch with retrieval-based methods that reuse text from similar past cases are more likely to succeed. Retrieval-based methods often work surprisingly well in DC as already discussed, a fact that DC research has not fully exploited yet, but some editing (or filling in) of recycled previous text (or templates) will presumably still be necessary in many cases, especially when reporting abnormalities. Hence, decoders that tailor (or fill in) previous diagnostic text (or templates) may still be needed. Reinforcement learning can be used to decide when to switch from recycling previous text to editing previous text or to generating new text, as already discussed. Ideally future work will also take into account that physicians do not consider only medical images when diagnosing. They also consider the medical history and profile of the patients (e.g., previous examinations, previous medications, age, sex, occupation). Hence, information of this kind (e.g., from electronic medical records) may need to be encoded, along

¹⁷<http://www.lire-project.net/>

with the images of the current examination, which may be more than one as already discussed. We also believe that DC methods may also need to become hybrid in the sense that they may need to involve more closely the physician who is responsible for a diagnosis. Current DC work seems to assume that systems should generate a complete diagnostic text on their own, which the responsible physician may then edit. In practice, however, it may be more desirable to allow the physician to see and correct regions of possible interest highlighted by the system on the medical images; then allow the physician to inspect and correct medical terms assigned by the system to the images; then let the physician start authoring the diagnostic text, with the system suggesting sentence completions, re-phrasings, missing sentences, in effect acting as an intelligent image-aware authoring tool. This would allow the physicians to monitor and control more closely the system’s predictions and decision making, especially if mechanisms to explain each system prediction or suggestion are available (e.g., highlighting regions on the images that justify each predicted term or suggested sentence completion).

Better intrinsic and extrinsic evaluation. We discussed the shortcomings of current automatic evaluation DC measures. Improving these measures to capture desirable properties of diagnostic text, especially clinical correctness, is hence an obvious area where further research is needed. Advances in evaluation measures for machine translation (Sun et al., 2020), summarization (Xenouleas et al., 2019) or, more generally, text generation (Sellam et al., 2020) also need to be monitored and ported to DC evaluation when appropriate. Despite the high cost, more manual evaluations of system-generated diagnostic reports by qualified physicians are also needed to obtain a better view of the real-life value of current DC methods and desired improvements. More extrinsic evaluations are also necessary, for example to check if DC methods can indeed shorten the time needed by a physician to author a diagnostic report, if DC methods indeed help inexperienced physicians avoid medical errors, if they reduce the pressure physicians feel etc. Extrinsic evaluations of this kind may also help shift DC methods towards hybrid forms that will involve physicians to a larger extent, as suggested above.

More, larger, realistic public datasets. We pointed out that there are currently only two datasets that are truly representative of the DC task. The first one, IU X-RAY, is rather small (approx. 4k instances) by today’s standards. The second one, MIMIC-CXR, is much larger (approx. 228k instances), but still small compared to the approx. 1 billion imaging examinations performed annually worldwide (Section 1), and it contains only English reports. Hospitals worldwide routinely save diagnostic medical images and the corresponding reports in their systems using established standards, at least for the images and their metadata (e.g., DICOM). Regulations and guidelines to protect sensitive information (e.g., HIPAA) are also available, and automatically removing sensitive information from both images and diagnostic reports seems feasible to a large extent (Section 2). Hence, it should be possible to construct many more and much larger publicly available DC datasets in many more languages; ideally these datasets would also include medical records and other information that physicians consult for diagnostic purposes, not just the medical images, as already discussed. What seems to be missing is a set of established, possibly standardized, procedures to construct publicly available and appropriately anonymized DC datasets. In turn, this requires concrete evidence (e.g., from extrinsic evaluations) of the possible benefits that DC methods may bring to public health systems, and well documented best practices.

Acknowledgements

We thank Vasiliki Liakopoulou for language editing of earlier drafts of this article.

References

- Aerts, H. J., Velazquez, E. R., Leijenaar, R. T., Parmar, C., Grossmann, P., Carvalho, S., Bussink, J., Monshouwer, R., Haibe-Kains, B., Rietveld, D., et al. (2014). Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications*, 5:1–9.
- Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., and Anderson, P. (2019). nocaps: novel object captioning at scale. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8948–8957, Seoul, Korea.
- Anderson, P., Fernando, B., Johnson, M., and Gould, S. (2016). SPICE: Semantic propositional image caption evaluation. In *Proceedings of the European Conference on Computer Vision*, pages 382–398, Amsterdam, Netherlands.
- Bai, S. and An, S. (2018). A survey on automatic image caption generation. *Neurocomputing*, 311:291–304.
- Banerjee, S. and Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine*

- Translation and/or Summarization of the Annual Conference of the Association for Computational Linguistics*, pages 65–72, Ann Arbor, MI, USA.
- Berlin, L. (2007). Accuracy of diagnostic procedures: Has it improved over the past five decades? *American Journal of Roentgenology*, 188:1173–1178.
- Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A., and Plank, B. (2016). Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55:409–442.
- Boag, W., Hsu, T.-M. H., McDermott, M., Berner, G., Alesentzer, E., and Szolovits, P. (2020). Baselines for chest x-ray report generation. In *Machine Learning for Health Workshop*, pages 126–140.
- Brady, A. P. (2017). Error and discrepancy in radiology: Inevitable or avoidable? *Insights into Imaging*, 8:171–182.
- Bustos, A., Pertusa, A., Salinas, J.-M., and de la Iglesia-Vayá, M. (2020). Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797.
- Charniak, E. (2018). *Introduction to Deep Learning*. MIT Press.
- Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., and Zitnick, C. L. (2015). Microsoft COCO captions: Data collection and evaluation server. *arXiv:1504.00325*.
- Chen, Z., Song, Y., Chang, T.-H., and Wan, X. (2020). Generating radiology reports via memory-driven transformer. In *to appear in the Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
- Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734, Doha, Qatar.
- Chokshi, F. H., Hughes, D. R., Wang, J. M., Mullins, M. E., Hawkins, C. M., and Duszak Jr, R. (2015). Diagnostic radiology resident and fellow workloads: A 12-year longitudinal trend analysis using national medicare aggregate claims data. *Journal of the American College of Radiology*, 12:664–669.
- de Herrera, A. G. S., Eickhoff, C., Andrearzyk, V., and Müller, H. (2018). Overview of the ImageCLEF 2018 caption prediction tasks. In *Proceedings of the CEUR Workshop, CLEF2018 Working Notes*, Avignon, France.
- Demner-Fushman, D., Kohli, M. D., Rosenman, M. B., Shooshan, S. E., Rodriguez, L., Antani, S., Thoma, G. R., and McDonald, C. J. (2015). Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23:304–310.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186, Minneapolis, MN, USA.
- Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., and Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2625–2634, Boston, MA, USA.
- Eickhoff, C., Schwall, I., de Herrera, A. G. S., and Müller, H. (2017). Overview of ImageCLEFcaption 2017 - the image caption prediction and concept extraction tasks to understand biomedical images. In *Proceeding of the CEUR Workshop, CLEF2017 Working Notes*, Dublin, Ireland.
- Fazal, M. I., Patel, M. E., Tye, J., and Gupta, Y. (2018). The past, present and future role of artificial intelligence in imaging. *European Journal of Radiology*, 105:246–250.
- Fellbaum, C. (2012). WordNet. *The encyclopedia of applied linguistics*.
- Friedlin, F. J. and McDonald, C. J. (2008). A software tool for removing patient identifying information from clinical documents. *Journal of the American Medical Informatics Association*, 15:601–610.
- Gale, W., Oakden-Rayner, L., Carneiro, G., Bradley, A. P., and Palmer, L. J. (2018). Producing radiologist-quality reports for interpretable artificial intelligence. *arXiv:1806.00340*.
- Gatt, A. and Krahmer, E. (2018). Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.
- Goldberg, Y. (2017). *Neural Network Methods in Natural Language Processing*. Morgan and Claypool Publishers.
- Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT press.
- Graham, Y. (2015). Re-evaluating automatic summarization with BLEU and 192 shades of ROUGE. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 128–137, Lisbon, Portugal.

- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, Las Vegas, NV, USA.
- Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9:1735–1780.
- Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., and Aerts, H. J. (2018). Artificial intelligence in radiology. *Nature Reviews Cancer*, 18:500–510.
- Hossain, M., Sohel, F., Shiratuddin, M. F., and Laga, H. (2019). A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys*, 51:1–36.
- Huang, G., Liu, Z., Maaten, L. V. D., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, Honolulu, HI, USA.
- Huang, X., Yan, F., Xu, W., and Li, M. (2019). Multi-attention and incorporating background information model for chest X-ray image report generation. *IEEE Access*, 7:154808–154817.
- Iandola, F., Moskewicz, M., Karayev, S., Girshick, R., Darrell, T., and Keutzer, K. (2014). DenseNet: Implementing efficient ConvNet descriptor pyramids. *arXiv:1404.1869*.
- Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoor, 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*, pages 590–597, Honolulu, HI, USA.
- Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM International Conference on Multimedia*, pages 675–678, Orlando, FL, USA.
- Jing, B., Xie, P., and Xing, E. (2018). On the automatic generation of medical imaging reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2577–2586, Melbourne, Australia.
- Johnson, A. E., Pollard, T. J., Berkowitz, S., Greenbaum, N. R., Lungren, M. P., Deng, C.-Y., Mark, R. G., and Horng, S. (2019). MIMIC-CXR: A large publicly available database of labeled chest radiographs. *arXiv:1901.07042*.
- Karpathy, A. and Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, Boston, MA, USA.
- Kassner, N. and Schütze, H. (2020). Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, held on-line.
- Khandelwal, U., Levy, O., Jurafsky, D., Zettlemoyer, L., and Lewis, M. (2020). Generalization through memorization: Nearest neighbor language models. In *Proceedings of the International Conference on Learning Representations*, pages 1–20, held on-line.
- Kilickaya, M., Erdem, A., Ikizler-Cinbis, N., and Erdem, E. (2016). Re-evaluating automatic metrics for image captioning. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, pages 199–209, Valencia, Spain.
- Kisilev, P., Sason, E., Barkan, E., and Hashoul, S. (2016). Medical image captioning: Learning to describe medical image findings using multi-task-loss CNN. In *Proceedings of the 1st International Workshop on Deep Learning for Precision Medicine*, Riva del Garda, Italy.
- Kisilev, P., Walach, E., Barkan, E., Ophir, B., Alpert, S., and Hashoul, S. Y. (2015a). From medical image to automatic medical report generation. *IBM Journal of Research and Development*, 59:1–7.
- Kisilev, P., Walach, E., Hashoul, S. Y., Barkan, E., Ophir, B., and Alpert, S. (2015b). Semantic description of medical image findings: Structured learning approach. In *Proceedings of the British Machine Vision Conference*, pages 171.1–171.11, Swansea, UK.
- Kougia, V., Pavlopoulos, J., and Androutsopoulos, I. (2019). A Survey on Biomedical Image Captioning. In *Proceedings of the Workshop on Shortcomings in Vision and Language of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 26–36, Minneapolis, MN, USA.
- Krupinski, E. A. (2010). Current perspectives in medical image perception. *Attention, Perception, & Psychophysics*, 72:1205–1217.
- LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521:436–444.
- Li, Y., Liang, X., Hu, Z., and Xing, E. (2018). Hybrid retrieval-generation reinforced agent for medical image report generation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 1537–1547, Montreal, Canada.

- Li, Y., Liang, X., Hu, Z., and Xing, E. (2019). Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6666–6673, Honolulu, HI, USA.
- Liang, S., Li, X., Zhu, Y., Li, X., and Jiang, S. (2017). ISIA at the ImageCLEF 2017 image caption task. In *Proceedings of the CEUR Workshop, CLEF2017 Working Notes*, Dublin, Ireland.
- Liew, C. (2018). The future of radiology augmented with artificial intelligence: A strategy for success. *European Journal of Radiology*, 102:152–156.
- Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out of the Annual Conference of the Association for Computational Linguistics*, pages 74–81, Barcelona, Spain.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755, Zurich, Switzerland.
- Liu, G., Hsu, T.-M. H., McDermott, M., Boag, W., Weng, W.-H., Szolovits, P., and Ghassemi, M. (2019a). Clinically accurate chest X-ray report generation. In *Proceedings of the Machine Learning for Healthcare Conference*, pages 1–20, Ann Arbor, MI, USA.
- Liu, X., Xu, Q., and Wang, N. (2019b). A survey on deep neural network-based image captioning. *The Visual Computer*, 35:445–470.
- Lu, J., Xiong, C., Parikh, D., and 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*, pages 375–383, Honolulu, HI, USA.
- Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to information retrieval*. Cambridge university press.
- Monshi, M. M. A., Poon, J., and Chung, V. (2020). Deep learning in generating radiology reports: A survey. *Artificial Intelligence in Medicine*, 106:101878.
- Mork, J. G., Jimeno-Yepes, A., and Aronson, A. R. (2013). The NLM medical text indexer system for indexing biomedical literature. In *Proceedings of BioASQ*, Valencia, Spain.
- Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.
- Nezami, O. M., Dras, M., Wan, S., and Paris, C. (2020). Image captioning using facial expression and attention. *Journal of Artificial Intelligence Research*, 68:661–689.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Philadelphia, PA, USA.
- Pelka, O., Friedrich, C. M., de Herrera, A. G. S., and Müller, H. (2019). Overview of the ImageCLEFmed 2019 Concept Prediction Task. In *Proceedings of the CEUR Workshop, CLEF2019 Working Notes*, Lugano, Switzerland.
- Pelka, O., Friedrich, C. M., Garcia Seco de Herrera, A., and Müller, H. (2020). Overview of the imageclefmed 2020 concept prediction task: medical image understanding. In *Proceedings of the CEUR Workshop, CLEF2020 Working Notes*, Thessaloniki, Greece.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1:9.
- Reiter, E. and Dale, R. (2000). *Building Natural Language Generation Systems*. Cambridge University Press.
- Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., and Goel, V. (2017). Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7008–7024, Honolulu, HI, USA.
- Schlegl, T., Waldstein, S. M., Vogl, W.-D., Schmidt-Erfurth, U., and Langs, G. (2015). Predicting semantic descriptions from medical images with convolutional neural networks. In *Proceedings of the International Conference on Information Processing in Medical Imaging*, pages 437–448, Isle of Skye, UK.
- Sellam, T., Das, D., and Parikh, A. P. (2020). Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, held on-line.
- Sharma, P., Ding, N., Goodman, S., and Soricut, R. (2018). Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2556–2565, Melbourne, Australia.

- Shin, H.-C., Lu, L., Kim, L., Seff, A., Yao, J., and Summers, R. M. (2016a). Interleaved text/image deep mining on a large-scale radiology database for automated image interpretation. *JMLR*, 17:3729–3759.
- Shin, H.-C., Roberts, K., Lu, L., Demner-Fushman, D., Yao, J., and Summers, R. M. (2016b). Learning to read chest X-rays: Recurrent neural cascade model for automated image annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2497–2506, Las Vegas, NV, USA.
- Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*.
- Suetens, P. (2009). *Fundamentals of Medical Imaging*. Cambridge University Press.
- Sun, S., Guzmán, F., and Specia, L. (2020). Are we estimating or guesstimating translation quality? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6262–6267, held on-line.
- Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.
- Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, 240:1285–1293.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826.
- Tsochantaridis, I., Hofmann, T., Joachims, T., and Altun, Y. (2004). Support vector machine learning for interdependent and structured output spaces. In *Proceedings of the International Conference on Machine Learning*, pages 104–114, Banff, Alberta, Canada.
- Van Deemter, K., Krahmer, E., and Theune, M. (2005). Real versus template-based natural language generation: A false opposition? *Computational Linguistics*, 31:15—24.
- Varges, S., Bieler, H., Stede, M., Faulstich, L. C., Irsig, K., and Atalla, M. (2012). SemScribe: Natural language generation for medical reports. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 2674–2681, Istanbul, Turkey.
- Vedantam, R., Zitnick, Z. C. L., and Parikh, D. (2015). CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575, Boston, MA, USA.
- Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, Boston, MA, USA.
- Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2017). Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:652–663.
- Wang, X., Peng, Y., Lu, L., Lu, Z., and Summers, R. M. (2018). TieNet: Text-image embedding network for common thorax disease classification and reporting in chest X-rays. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9049–9058, Quebec City, Canada.
- Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256.
- Xenouleas, S., Malakasiotis, P., Apidianaki, M., and Androutsopoulos, I. (2019). Sumqe: a bert-based summary quality estimation model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 6005–6011, Hong Kong, China.
- Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning*, pages 2048–2057.
- Xue, Y., Xu, T., Long, L. R., Xue, Z., Antani, S., Thoma, G. R., and Huang, X. (2018). Multimodal recurrent model with attention for automated radiology report generation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 457–466, Granada, Spain.
- Yin, C., Qian, B., Wei, J., Li, X., Zhang, X., Li, Y., and Zheng, Q. (2019). Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network. In *Proceedings of the IEEE International Conference on Data Mining*, pages 728–737, Beijing, China.
- You, Q., Jin, H., Wang, Z., Fang, C., and Luo, J. (2016). Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4651–4659, Las Vegas, NV, USA.

- Yuan, J., Liao, H., Luo, R., and Luo, J. (2019). Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 721–729, Shenzhen, China.
- Zhang, Y., Merck, D., Tsai, E. B., Manning, C. D., and Langlotz, C. P. (2019). Optimizing the factual correctness of a summary: A study of summarizing radiology reports. *arXiv:1911.02541*.
- Zhang, Y., Wang, X., Guo, Z., and Li, J. (2018). ImageSem at ImageCLEF 2018 caption task: Image retrieval and transfer learning. In *Proceedings of the CEUR Workshop, CLEF2018 Working Notes*, Avignon, France.
- Zhang, Z., Chen, P., Sapkota, M., and Yang, L. (2017a). TandemNet: Distilling knowledge from medical images using diagnostic reports as optional semantic references. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 320–328, Quebec City, Canada.
- Zhang, Z., Xie, Y., Xing, F., McGough, M., and Yang, L. (2017b). MDNet: A semantically and visually interpretable medical image diagnosis network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6428–6436, Honolulu, HI, USA.