Cogent Engineering

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/oaen20

Semantic interdisciplinary evaluation of image captioning models

Uddagiri Sirisha & Bolem Sai Chandana

To cite this article: Uddagiri Sirisha & Bolem Sai Chandana (2022) Semantic interdisciplinary evaluation of image captioning models, Cogent Engineering, 9:1, 2104333, DOI: 10.1080/23311916.2022.2104333

To link to this article: https://doi.org/10.1080/23311916.2022.2104333

Published online: 21 Aug 2022.

Sirisha & Sai Chandana, *Cogent Engineering* (2022), 9: 2104333
https://doi.org/10.1080/23311916.2022.2104333

cogent engineering

COMPUTER SCIENCE | REVIEW ARTICLE

Semantic interdisciplinary evaluation of image captioning models

Uddagiri Sirisha[1]* and Bolem Sai Chandana[1]

Received: 26 March 2022
Accepted: 21 June 2022

*Corresponding author: Uddagiri Sirisha, School of Computer, Science & Engineering(SCOPE), VIT-AP University, Amaravathi, 52223, Andhra Pradesh, India
E-mail: sirisha.uddagiri@gmail.com

Reviewing editor:
Marko Robnik-Šikonja, Faculty of Computer and Information Science, University of Ljubljani, Slovenia

Additional information is available at the end of the article

Abstract: In our day-to-day life, synchronizing vision and language aspects plays a crucial role in solving various real-time challenges. Image captioning is one of them, and it aims to recognise objects, activities, and their relationships in order to provide a syntactically and semantically correct visual description. There are existing works of image captioning in various directions, such as news, fashion, art, and medical domains. The core architectural idea of image captioning is based on merging CNN, RNN, and transformer models. In practice, there are many conceivable combinations, and brute forcing all of them would take a long time. As we know, there is no work on interpreting image captioning models across various usecases. In this research article, we examine and analyze different image captioning models used across various domains, and multiple insights are extracted to determine the best combinational architecture for a new application without ignoring contextual semantics. We examined numerous designs and determined that LSTM is best for image captioning across several domains.

Subjects: Cognitive Artificial Intelligence; Neural Networks; Computer Science; General

Keywords: image captioning; news articles; interdisciplinary; fashion images; visually impaired people; medical images; art images; evaluation metrics

1. Introduction

Visual data like photographs and videos may now be acquired fast and inexpensively, providing a wealth of knowledge for addressing real-world problems. The availability of huge amounts of

ABOUT THE AUTHOR

Uddagiri Sirisha, is a PhD full time research scholar in the school of Computer Science and Engineering from VIT-AP University, Amaravati. The author was awarded B.Tech and M.Tech degree from JNTU Kakinada and had 5 years of teaching experience. Her research interests include Image Processing, Deep Learning, Machine Learning.
Dr.B.Sai Chandana, working as Assistant Professor Sr.Grade 2 in the school of Computer Science and Engineering. Her research interests include hyperspectral image processing, remote sensing, medical image analysis, and activity recognition. She can be reached at saichandana.bolem@vitap.ac.in She published 30 research papers out of which 15 are included in the SCOPUS database. She published three articles in CSI communications.

PUBLIC INTEREST STATEMENT

Image captioning has various applications such as recommendations in editing applications, usage in virtual assistants, for image indexing, for visually impaired persons, for social media, and several other natural language processing applications. Many researchers have been doing research in this field of image captioning using deep learning architectures. The paper contains brief study for the methods involved in a variety of contexts, including news articles, fashion, medical photos, art images, and even human images.

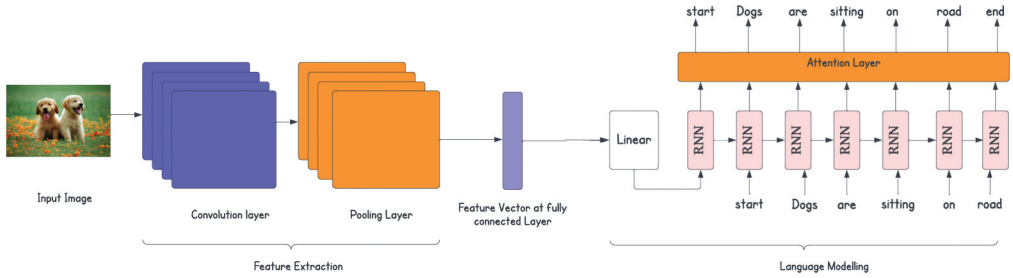


Figure 1. Image caption generator using CNN & LSTM.

data creates demands for automatic visual understanding and content summarization, which is not possible in real time. Humans can understand themselves without a description of an image, and models will not inherit this flexibility until an appropriate hybrid architectural combination is established. As a result, image captioning is used to describe images by using various deep learning models.

Captioning is a tool for describing the content of an image by utilising computer vision and natural language processing. Captioning allows the model to recognise not only the objects in an image but also their relationships with other objects to generate human-readable phrases.

Captions are generated by using a CNN-RNN architecture, which involves CNN layers for feature extraction on input data and RNN to make predictions based on time series data. A sample diagram is shown in Figure 1.

A convolutional neural network is an algorithm that takes an input image and applies weights and biases to different items in the image to distinguish one image from another. Many convolutional layers are integrated with nonlinear and pooling layers in each neural network. When a picture is streamed through each convolution layer, the first layer's output becomes the input for the next layer, and so on for all subsequent layers. To construct an N-dimensional vector, a fully connected layer is required after a series of convolutional, nonlinear, and pooling layers.

Long short-term memory networks are a sort of recurrent neural network that can learn order dependence in sequence prediction problems and overcome the RNN's short-term memory constraints. LSTM can save useful information throughout the processing of inputs while discarding irrelevant data. The prior inputs have a longer reference in LSTM and GRU, but only to a limited extent. To overcome this, we use transformers to provide an unlimited reference to the prior inputs.

A transformer is an encoder and decoder architecture that transforms one sequence into another using only attention. The encoder transforms an input sequence into an abstract continuous representation that includes all the data. The decoder then takes the encoder's continuous representation and produces a single output, which is then fed back into the preceding output.

The domains listed below are used for image captioning:

1) Hand crafted features are learned using techniques like Scale-Invariant Feature Transform (SIFT; Lowe, 2004), Local Binary Patterns (LBP; Lowe, 2004), the Histogram of Oriented Gradients (HOG; Dalal & Triggs, 2005), and a combination of these approaches. To classify an object, extracted features are fed into a classifier like a support vector machine (SVM; Boser et al., n.d.). As handcrafted features are job dependent, extracting features from a huge and diverse amount of data is tough. To classify an object, extracted features are fed into a classifier like a support vector machine (SVM; Boser et al., n.d.). As handcrafted features are job-dependent, extracting features from a huge and diverse amount of data is tough.

2) Deep-learning-based methods are automated to classify both linear and non-linear feature spaces.

The following methods are used for deep learning-based image captioning (Adriyendi, 2021):

Template/Retrieval-based methods:Different objects, properties, actions, and their relationships are identified using a template-based technique. Then the empty space in the templates is filled. From the training data, the retrieval-based approach detects visually similar images and then generates captions. A novel approach is used to create captions from both the visual and multi-modal spaces.

Dense/Single Sentence Captioning: Dense captioning (Johnson et al., 2016) generates captions for each region in the image, whereas single sentence captioning generates captions for the entire scene in the image.

Architecture/Compositional-based image captioning: Architecture based image captioning involves convolutional neural network and recurrent neural network (Vinyals et al., 2015) whereas compositional-based approaches include (Fang et al., 2015), GAN based (C. Chen et al., 2019), Semantic concept-based (You et al., 2016), Graph based (Pan et al., 2004), Attention-based (Xu et al., 2015), and stylized captions (Gan et al., 2017). In encoder-decoder architecture-based image captioning, the visual features are extracted using CNN approach and fed into the LSTM for the caption generation. But in compositional architecture-based image captioning, the visual features and also the attributes are extracted using CNN approaches and fed into the language models for the caption generation. They are then re-ranked to identify high-quality image captions using a deep multimodal architecture.

In recent years, there has been a decent work for image captioning because of the capability of deep learning models to efficiently extract useful features from images. Most of the deep learning models use convolution neural networks (LeCun et al., 1998) for extracting image features and language models for caption generation of an image. The most common CNN approaches are Resnet-50, VGG-16, Inception v3, Densenet, FasterR-CNN, etc., and the most frequently used RNN models are LSTM and its successors i.e. single-layer LSTM, single-layer LSTM with attention, Bidirectional LSTM, Hierarchical LSTM, GRU and Transformers.

Common datasets involved in image captioning are MS COCO (Lin et al., 2014), Flickr30K (Plummer et al., 2015), Flickr8K (Hodosh et al., 2013), Visual Genome (Krishna et al., 2017), Instagram (Chunseong Park et al., 2017; K. Tran et al., 2016), IAPR TC-12 (Grubinger et al., 2006), Stock3M, MIT-Adobe FiveK (Bychkovsky et al., 2011), FlickrStyle10k.

Model performance depends on two categories of metrics.

1) To measure the overall compatability between generated captions and ground truth.

2) Calculation of precision and recall scores.

The most commonly used evaluation metrics for image captioning include BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), METEOR (Banerjee & Lavie, 2005), CIDER (Vedantam et al., 2015), SPICE (Anderson et al., 2016). The importance of each metric is mentioned below:

1.1. BLEU

BLEU stands for Bilingual Evaluation Understudy Score and is initially proposed for machine translation. BLEU is a metric to measure the compatibility between two text strings. Bleu is applied on the test dataset.

The Bleu score is determined as follows:

1) Let us consider a phrase s, a list of possible reference phrases and a candidate phrase c.

2) Initially, we calculate the modified precision p_n of c where n = 1,2,3 … .n.

$$p_n = \frac{No\ of\ times\ N_c\ appears\ in\ S_r}{No\ of\ n-grams\ in\ S_c}$$

where N_c represents the n-grams in candidate phrase, S_r represents the reference phrase and S_c represents the candidate phrase.

3) Recall is a measure of quantity and it measures the entire content of the output whereas precision is a measure of quality as it measures the n-grams score separately. Recall is not considered in BLEU metric and so in order to satisfy for recall, BLEU uses a brevity penalty.

4) Let S *r* be the reference phrase length and S_c be the candidate phrase length. We compute the brevity penalty(P) as follows

$$P = \begin{pmatrix} 1, & \text{if } S_c > S_r; \\ e^{1-S_r/S_c}, & \text{if } S_c \leq S_r; \end{pmatrix}$$

5) The geometric mean of the n-gram precision score multiplied by the shortness penalty(P) for short phrases generates BLEU score. Then, the BLEU score is

$$BLEU = \mathrm{P.exp}\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

6) The BLEU metric goes from 0 to 1. To get a score of 1, the phrase must be exactly like the reference phrase. Otherwise, the score is between 0.1 and 0.9%. As a result, even a human translator will not always get a perfect score of 1.

1.2. METEOR

METEOR is one of the evaluation metric for machine translation. METEOR addresses the disadvantages of BLEU metric such as recall evaluation, absence of explicit word matching. Based on word-to-word similarities between a candidate and a reference phrase, meteor ratings are generated. If more than one reference phrase is provided, the score is calculated individually against each reference. Meteor chooses the alignment with the most similar word order between the two strings. To discover the best alignment between two strings, the exact mapping module is utilised first, followed by stemming and synonyms. Later, we calculate unigram precision p = m / c, recall r = m / r, where m denotes the mapped unigrams between the two strings and c,s denotes the total number of words in candidate, reference phrases, respectively.

The harmonic mean of precision and recall of unigram matching between candidate and reference phrases computes METEOR score. The harmonic mean is given as:

$$mean = \frac{10pr}{r+9p}$$

METEOR groups the unigrams in the candidate and reference phrases into chunks, i.e. when the entire candidate phrase matches the reference phrase then there is only one chunk and if there is

no match, then there are many chunks. Then the penalty for the chunks is computed using the formula:

$$penalty = 0.5 * \left(\frac{chunks}{unigrams - matched}\right)^3$$

For each additional chunk, penalties increase by a maximum of 0.5% and then drop to their minimum value based on the number of matching unigrams. Finally, the metor score M_s is calculated as:

M_s= mean*(1-penalty)

1.3. SPICE

SPICE stands for semantic propositional image phrase evaluation. SPICE is used for finding image caption similarity using a scene graph. Therefore, in SPICE, we transform both candidate phrase and reference phrases into an intermediate representation such as scene graph.

For a given candidate phrase c and a set of reference phrases s = $(s_1, s_2, \ldots, s_m)$ associated with an image, it computes similarity between c and s. An object class (C), a relation (R), and an attribute (A) will be parsed from the given phrase utilising the scene graph. In order to create a scene graph, the candidate phrase c is first transformed into a logical proposition i.e.

SG_c=< O_c,E_c,K_c>–where SG_c denotes scene graph of phrase c.

$O_c \subseteq$ C—where O(c) denotes a set of objects.

$E_c \subseteq O_c$ X R X O_c—where E(c) denotes set of hyper edges.

$K_c \subseteq O_c$ X A—where K(c) denotes set of attributes.

The performance of SPICE depends on parsing and its value is bounded between 0 and 1.

1.4. CIDER

CIDER stands for consensus-based image description evaluation. It compares the consistency of a candidate phrase to a set of human-written reference phrases. CIDEr, on the other hand, has a higher agreement with consensus as measured by humans. The characteristics of saliency, grammaticality, semantics and accuracy are essentially captured by this measure.

For a given candidate phrase c and a set of reference phrases s = $(s_1, s_2, \ldots, s_m)$ associated with an image, initial stemming is applied and each phrase is represented in the form of n-grams. The candidate and reference phrases n-gram co-occurrences are then calculated. The n-grams that appear in all image descriptions are given a lower weight in the CIDER metric. The cosine similarity between the candidate and the reference n-grams can be determined after performing a term frequency inverse document frequency (TF-IDF) weighting on each n-gram.

1.5. ROUGE

ROUGE means recall-oriented understudy for gisting evaluation, It was created with the intention of assessing summarization systems. The amount of overlapping units, such as n-grams, word sequences, and word pairs, is measured by ROUGE. We have four categories of ROUGE, i.e. ROUGE-L: which basically measures the Longest Common Subsequence, ROUGE-N: N-gram Co-Occurrence Statistics, ROUGE-W: Weighted Longest Common Subsequence, ROUGE-S: Skip-Bigram Co-Occurrence Statistics. ROUGE metric relies highly on recall, it favors long phrases.

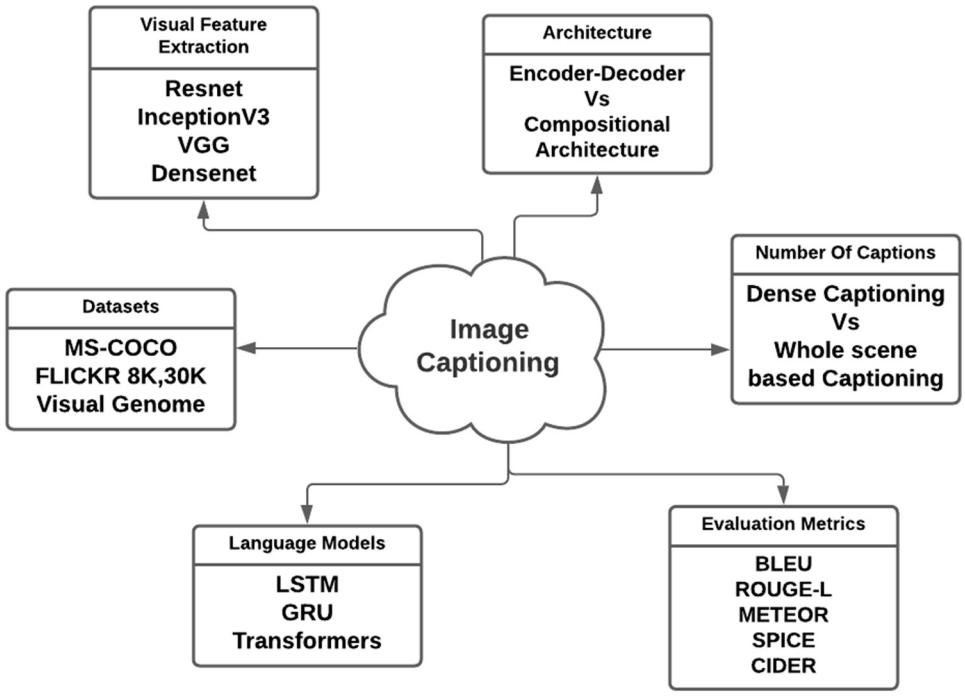


Figure 2. Image captioning possible candidates.

The entities which are part of image captioning life cycle are shown in Figure 2.

2. Related work

2.1. Prior analysis in image captioning

A huge number of articles on image captioning have been published in recent years. As summarised in Table 1, only a few survey studies have been published, each of which gave a decent literature review of image captioning. The writers of (Hossain et al., 2019) looked at deep learning-based image captioning algorithms, a taxonomy of image captioning techniques, various

Table 1. Comparative analysis on image captioning surveys

Reference	Template/ Retrieval based methods	Architectures	Datasets	Evaluation metrics	Interdisciplinary domain
(Hossain et al., 2019)	YES	YES	YES	YES	NO
(Stefanini et al., n.d.)	NO	YES	YES	YES	NO
(Choi et al., 2021)	NO	YES	NO	NO	NO
(Elhagry & Kadaoui, 2021)	YES	YES	YES	YES	NO
(Oluwasammi et al., 2021)	YES	YES	YES	YES	NO
(Y. Wang et al., 2019)	YES	YES	YES	YES	NO
(Staniūtė & Šešok, 2019)	YES	YES	YES	YES	NO
(Bai & An, 2018)	YES	YES	YES	NO	NO
(X. Liu et al., 2019)	YES	YES	NO	YES	NO
(Pavlopoulos et al., 2021)	YES	YES	YES	YES	NO
(Monshi et al., 2020)	YES	YES	YES	YES	NO
Our Paper	YES	YES	YES	YES	YES

assessment criteria, and datasets. The authors of (Stefanini et al., n.d.) presented a literature review and experimental comparison on standard datasets w.r.t to performance, Choi et al. (2021) performed a comparative analysis of feature extraction methods in terms of image captioning.

Updown, OSCAR, VIVO, Meta Learning, and GAN-based models are among the state-of-the-art techniques examined by the authors in (Elhagry & Kadaoui, 2021). The advancements in semantic segmentation were discussed by the authors in (Oluwasammi et al., 2021). Template-based, retrieval-based approaches and as well as current improvements in encoder-decoder structure were discussed by the authors Y. Wang et al. (2019), X. Liu et al. (2019), and Bai & An (2018).

In Monshi et al. (2020), the authors presented a literature survey on multi-modal datasets for training deep DL models that generate radiology text. Pavlopoulos et al. (2021) presented an overview of publicly available datasets, evaluation measures, and encoder decoder architectures in the medical domain.

The survey papers (Bai & An, 2018; Choi et al., 2021; Elhagry & Kadaoui, 2021; Hossain et al., 2019; X. Liu et al., 2019; Monshi et al., 2020; Oluwasammi et al., 2021; Pavlopoulos et al., 2021; Staniūtė & Šešok, 2019; Stefanini et al., n.d.; Y. Wang et al., 2019) mainly discussed template, retrieval-based image caption models, different CNN/RNN architectures, different datasets, different evaluation metrics, etc, but could not showcase any interdisciplinary pattern interpretations of image captioning models. We fill this gap by presenting current research in a more thorough manner, as well as adding our explanations to cross domain image captioning models.

2.2. Novelty and contributions

The novelty and contributions of this work include

• We investigate the performance of various image captioning algorithms in a variety of contexts, including news articles, fashion, medical photos, art images, and even human images.

- News image captioning creates meaningful captions for photos from news articles that have been published in a variety of newspapers.

- High-level semantic information, professional knowledge, and several specific symbols are frequently included in art image captioning.

- Medical captioning encodes and generates a report from medical reports taken during a patient's examination.

- Image captioning helps visually impaired people to navigate more easily.

• From diverse domains, we developed a list of image captioning models. However, there is a lack of understanding of the model's performance in response to a variety of parameters. We apply statistical analysis to overcome this problem and reduce the amount of time spent on brute forcing for new applications.

3. Image captioning in various domains

The main theme of the paper is to reduce the time in finding out the best possible CNN and RNN architectures irrespective of the domain under consideration. Various domains and organization of this section are shown in Figure 3.

3.1. News image captioning

News image captioning differs from generic image captioning in three aspects:

1) Input consists of image-article pairs.

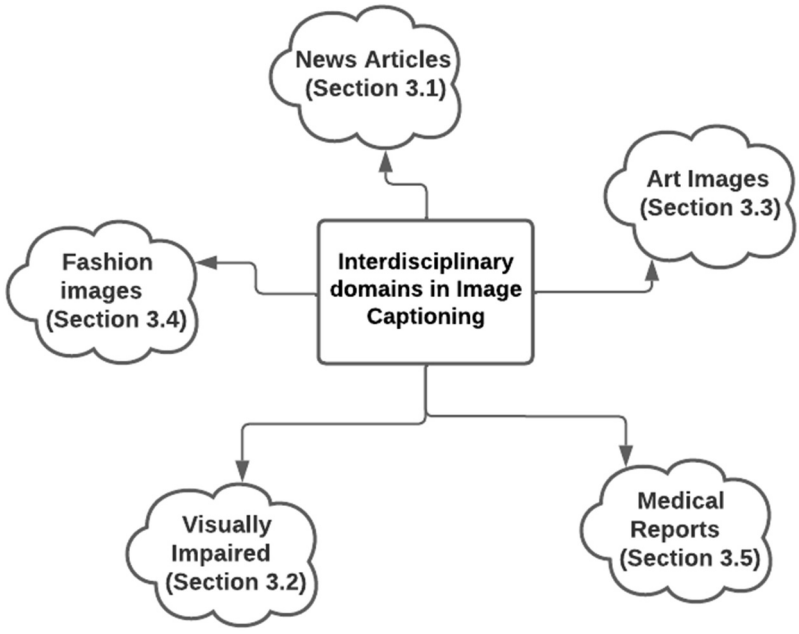


Figure 3. Interdisciplinary domains of image captioning.

2) Caption generation is performed based on the image and its related article.

3) News captions contain named entities and additional context extracted from the article.

Considering these differences, a sample image caption on news article is shown in Figure 4.

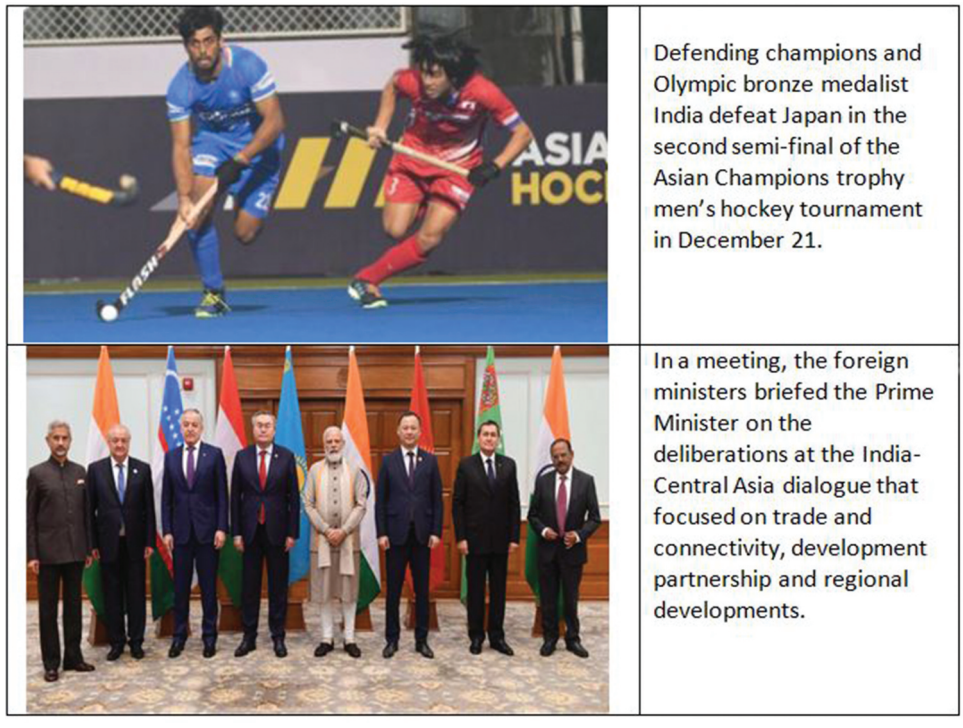


Figure 4. News image captioning.

Table 2. Datasets in news image captioning

Dataset	Number of samples	Description
Visual News (Gurari et al., 2019)	1080 K	Visual News dataset collected from a diverse set of news sources such as the Guardian, BBC, USA TODAY, and The Washington Post.
GoodNews (Furkan Biten et al., 2019)	466 K	Good News dataset collected from the New York Times news articles ranging from the year 2010 to 2018.
NYTimes800K (A. Tran et al., 2020)	792 K	NYTimes800K dataset collected from the New York Times news articles. In NYTimes800k, 97 percent of captions contains at least one named entity.
Breaking News (Ramisa et al., 2017)	115 K	The Breaking News dataset contains over 100,000 items, all of which contain at least one image and these include news about sports, politics, and health care.
BBC News	3 K	A document, an image, and its caption can all be found in the BBC News database. The dataset includes information on national and international politics, technology, sports, education, and other issues.
Daily Mail	209 K	DailyMail corpora news articles are collected from the DailyMail news websites.

News image captioning is used in editing applications. News image captioning is the process of creating detailed and informative captions for photos from news articles published in various newspapers. News article contains the objects along with the text and they refer to various locations, people, events, time, etc. So, the captions generated depict the visual and textual features of a news article. The most commonly used datasets are Visual News, Good News, NYTimes800K, Breaking News, BBC News, Daily Mail dataset. Table 2 shows various datasets used in news image captioning.

Analysis of architectures used in news image captioning is shown in Table 3. In state-of-the-art approaches, the authors (Hu et al., 2020; Liu et al., 2020; Yang et al., 2021; Zhao et al., 2021) specifically used attention-based encoder-decoder frameworks for generating captions for news article images. Mostly, in news image captioning the authors (Furkan Biten et al., 2019; Hu et al., 2020; Liu et al., 2020; A. Tran et al., 2020; Yang et al., 2021; Yang & Okazaki, 2020; Zhao et al., 2021) used RESNET-152, Chen and Zhuge (2019); Ramisa et al. (2017) applied Oxford VGGNET for visual feature extraction.

Named entity recognition is critical for identifying and classifying named entities in news image captioning. The identified entities are divided into numerous categories based on the raw and structured text such as persons, organisations, places, money, time, and so on. The writers of (Hu et al., 2020; Liu et al., 2020) utilized spaCy toolkit, (Yang et al., 2021) applied named entity embedding (NEE), and multi-span text reading (MSTR) for captioning.

Text feature extraction from news stories can be done at the article or sentence level, however it is not enough to collect all of the information in the article. For text classification, (Liu et al., 2020) applied word embedding and position embedding, A. Tran et al. (2020), Yang et al. (2021), and Zhao et al. (2021) used RoBERTa, Chen and Zhuge (2019); Rane et al. (2021) utilized word2Vec embeddings

Table 3. Analysis of various parameters used in news image captioning

Reference	Image extractor	Text extractor	Datasets	Evaluation metrics
Liu et al. (2020)	Resnet-152	Word Embedding and Position Embedding +spaCy	Visual News, GoodNews, NYTimes800K	BLEU-4, ROUGE, METEOR, CIDER.
Yang et al. (2021)	Resnet-152	RoBERTa+Named Entity Embedder+ multi-span text reading	GoodNews, NYTimes800K	BLEU-4, ROUGE, METEOR.
Hu et al. (2020)	Resnet-152	Glove word embeddings+ SpaCy	Breaking News, GoodNews	BLEU-4, ROUGE, METEOR, CIDER.
Zhao et al. (2021)	ResNet-152	RoBERTa	GoodNews, NYTimes800K	BLEU-4, ROUGE, METEOR, CIDER.
A. Tran et al. (2020)	Resnet-152	RoBERTa+Byte pair Encodings	GoodNews, NYTimes800K	BLEU-4, ROUGE, METEOR.
Yang and Okazaki (2020)	ImageNet	Glove word embeddings	GoodNews	BLEU, ROUGE, METEOR, CIDER, SPICE.
Furkan Biten et al. (2019)	Resnet-152	Glove word embeddings	GoodNews	BLEU-1, BLEU-2, BLEU-3,BLEU-4, ROUGE, METEOR, CIDER,SPICE.
Chen and Zhuge (2019)	Oxford VGGNET	Word2Vec embeddings	Daily Mail	BLEU, ROUGE, METEOR.
Batra et al. (2018)	Oxford VGGNET	Order Embeddings	BBC News dataset	BLEU, METEOR.
Ramisa et al. (2017)	VGG19	Word2Vec embeddings	Breaking News	BLEU-1, BLEU-2, BLEU-3,BLEU-4, METEOR.

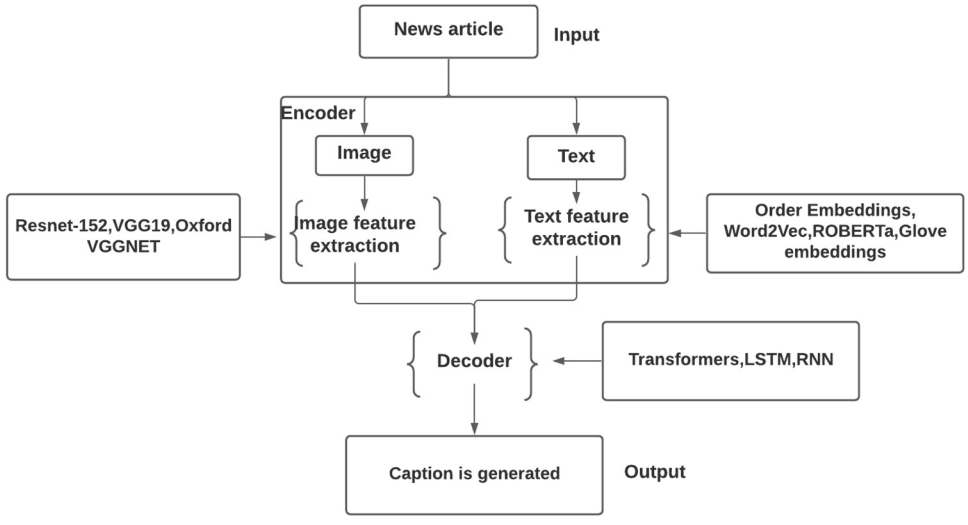


Figure 5. Workflow of news image captioning.

and Furkan Biten et al. (2019); Yang and Okazaki (2020) used glove word embeddings. The authors Hu et al. (2020), A. Tran et al. (2020) uses LSTM-BPE, sentence level approach in their models.

Good News(Furkan Biten et al., 2019; Hu et al., 2020; Liu et al., 2020; A. Tran et al., 2020; Yang et al., 2021; Zhao et al., 2021) and NYTimes800K (Liu et al., 2020; A. Tran et al., 2020; Yang et al., 2021; Zhao et al., 2021) are the most widely used datasets in news image captioning. The authors

Table 4. Performance summarization in news image captioning

Reference	Datasets	BLEU	METEOR	ROUGE	CIDER	SPICE
Liu et al. (2020)	Visual News	5.3 (BLEU-4)	8.2	17.9	50.5	-
	GoodNews	6.1 (BLEU-4)	8.3	21.6	55.4	-
	NYTimes800K	6.4 (BLEU-4)	8.1	21.9	56.1	-
Yang et al. (2021)	GoodNews	6.83 (BLEU-4)	23.05	11.25	61.22	-
	NYTimes800K	6.79 (BLEU-4)	22.80	10.93	59.42	-
Hu et al. (2020)	Breaking News	1.71 (BLEU-4)	6.00	14.33	25.74	-
	GoodNews	1.96 (BLEU-4)	6.01	15.70	26.08	-
Zhao et al. (2021)	GoodNews	5.89 (BLEU-4)	5.98	19.19	45.22	-
	NYTimes800K	5.97 (BLEU-4)	6.03	19.20	47.27	-
A. Tran et al. (2020)	GoodNews	6.05 (BLEU-4)	-	21.4	54.3	-
	NYTimes800K	6.30(BLEU-4)	-	21.7	54.4	-
Yang and Okazaki (2020)	GoodNews	8.78(BLEU-1) 10.9 (BLEU-2) 6.76 (BLEU-3) 4.52 (BLEU-4)	8.62	20.56	44.16	10.19
Furkan Biten et al. (2019)	GoodNews	1.60(BLEU-1) 3.54 (BLEU-2) 1.60 (BLEU-3) 0.83 (BLEU-4)	4.34	8.92	12.79	4.25
Chen and Zhuge (2019)	Daily Mail	7.24(BLEU)	16.68	26.07	-	-
Ramisa et al. (2017)	BBC News dataset	0.3427(BLEU)	0.0706	-	-	-
Rane et al. (2021)	Breaking News	19.6(BLEU-1) 8.9 (BLEU-2) 3.4(BLEU-3) 1.9(BLEU-4)	5.3	-	-	-

in (Liu et al., 2020) introduced the largest visual news image captioning dataset consisting of one million images with articles, captions, and other metadata. The model performance is effective on visual news dataset when compared to other two datasets, i.e. GoodNews and NYTimes800K.

A detailed workflow of the process involved in news image captioning is shown in Figure 5

• The input for image captioning is a news article {image +Text} and they need to be processed separately.

• The image encoder uses multiple CNN models to extract characteristics from the image. Various embedding approaches are used by text encoder to extract the features from news articles.

• Decoder combines both the visual and textual features using various RNN models and generates the caption as output.

News image captioning evaluation metrics analysis is shown in Table 4. Among all metrics, CIDER is the best performance metric in news captioning since it has dedicated embedding feature to focus more on un-common words. The authors in (Liu et al., 2020) introduced a new dataset, i.e. visual news dataset and the proposed model is able to generate better captions in a more efficient way, i.e. from 13.2 to 50.5 in CIDER score. The authors in (Zhao et al., 2021) constructed a multi-modal knowledge graph, and the authors in (Yang et al., 2021) even integrated domain-specific knowledge and achieved the best results on both datasets, i.e. GoodNews dataset and the NYTimes800k dataset.

Table 5. Datasets used in image captioning for visually impaired

Dataset	Number of samples	Description
VizWiz(Gurari et al., 2019)	39 K	(Gurari et al., 2019) The VizWiz dataset contains almost 39,000 images created by blind persons, each of which has five captions.
Flickr8K (Hodosh et al., 2013)	8 K	Flickr8K consists of 8092 and the images were chosen from six different Flickr groups.
MsCOCO(Lin et al., 2014)	328 K	The MS COCO dataset contains 328 K images, each of which is accompanied by five captions.

Table 6. Analysis of various parameters used in image captioning for visually impaired people

Reference	CNN/RNN	Deployment endpoint	Datasets	Evaluation metrics
Rane et al. (2021)	InceptionV3/LSTM	Google Text-to-Speech	Flick8k	BLEU-4.
Dognin et al. (2020)	ResNeXT/LSTM	Watson Text to Speech	VizWiz	BLEU, METEOR, ROUGE, CIDER, SPICE.
Pasupuleti et al. (2021)	VGG16/Guided LSTM	Google Text-to-Speech	Flick8k	BLEU-1, BLEU-2, BLEU-3,BLEU-4.
Ahsan et al. (2021)	AOANET/LSTM	-	VizWiz	BLEU, ROUGE, CIDER, SPICE .
Chharia and Upadhyay (2020)	Impaired VGG16/ LSTM	Anaroid application	Flick8k	BLEU .
Makav and Kılıç (2019a)	VGG16/ LSTM	Anaroid Application	MSCOCO	BLEU-1, BLEU-2, BLEU-3,BLEU-4, CIDER.
Makav and Kılıç (2019b)	VGG16/ LSTM	Anaroid application	Flick8k, MSCOCO	-
Zaman et al. (2019)	VGG16/ LSTM	text letters to braille	Flick8k	BLEU4.

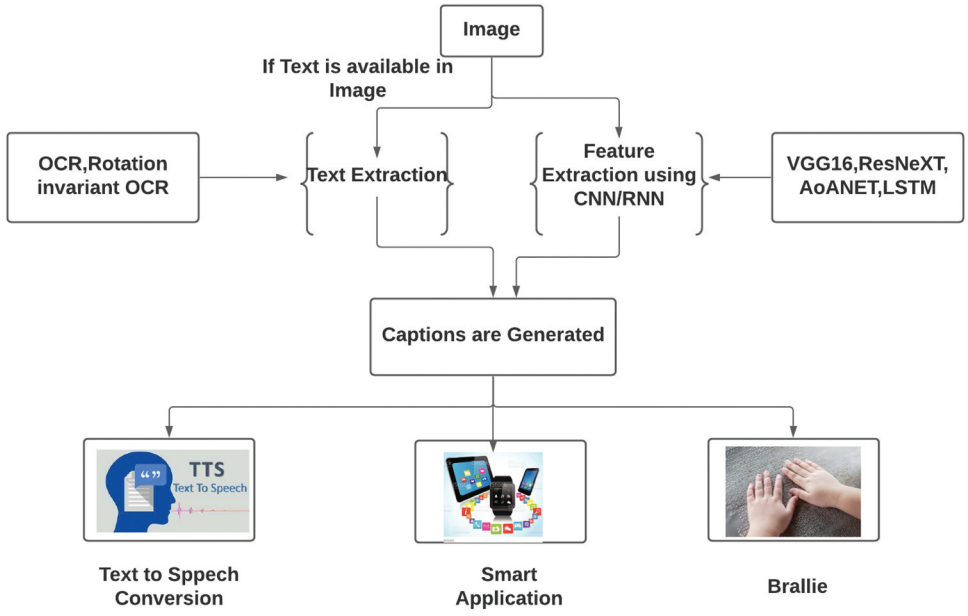
News image captioning is an inherently complicated challenge for machine intelligence, because it combines images and news articles for caption development. A detailed workflow of the process involved in news image captioning is also outlined. We summarized the work across various parameters used in news image captioning. News image captioning architectures are further evaluated across different datasets.

3.2. Image captioning for visually impaired

Visually impaired people face a number of challenges every day—from reading the label on a frozen dinner to figuring out ways to reach home safely. These problems will get worsen when they travel to new places. Many tools based on computer vision and other sensors have been developed to address these issues (talking OCR, GPS, radar canes, etc.). But adding a different perspective to solve these problems is possible with deep learning. Various data science researchers addressed several problems for visually impaired and one of them is generating image/scene captions considering the capabilities of respective challenged person. Image captioning helps visually impaired people to navigate more easily. The most commonly used datasets are VizWiz dataset, Flick8k, MSCOCO (Table 5).

To create captions for the given input image, state-of-the-art algorithms primarily focused on image captioning models. To create caption, an input image is passed through CNN to extract

Figure 6. Workflow of image captioning for visually impaired people.



visual features, and the outputs are merged and submitted to a multimodal transformer network. Various image captioning architectures used for visually impaired people are shown in Table 6.

In (Dognin et al., 2020), the authors developed a multi-modal transformer that uses ResNext visual features, object detection-based textual features, and OCR-based textual features. The authors in (Ahsan et al., 2021) used AoANet as a captioning model and BERT to build OCR token embeddings. In (Makav & Kılıç, 2019b; Pasupuleti et al., 2021; Zaman et al., 2019), the authors utilized VGG16 for feature extraction and Guided LSTM for text generation, Makav and Kılıç (2019a) applied VGG16 for visual features and NLP model for generating human-like captions.

The models build in this domain are deployed in edge and IoT devices. For example, the authors integrated the developed model into (Rane et al., 2021), (Makav & Kılıç, 2019a) into smart phones, (Chharia & Upadhyay, 2020) converted the captions to audio, (Zaman et al., 2019) integrated the model to be available in Braille. The authors in (Makav & Kılıç, 2019b) designed an android application "Eye of Horus" to provide an user-friendly interface, i.e. user can either choose an image from the gallery or take a new photo using the smartphone camera to produce captions. The caption can be listened and also displayed on the screen.

We came up with a common workflow based on the all the existing systems (Figure 6.) and the same is briefed as below:

- The input for image captioning is an image.

- Image features need to be extracted using CNN model and sent to RNN model for further processing.

- The model can even extract text from image using state-of-the-art OCR.

- Later, the results obtained from feature extraction and text extraction are combined to generate a caption.

- The generated caption or model can be used in the form of voice using GTTS, deployed in any smart device and also it can be converted to Braille.

Table 7. Performance summarization in image captioning for visually impaired people

Reference	Datasets	BLEU	METEOR	ROUGE	CIDER	SPICE
Rane et al. (2021)	Flick8k	0.24(B-4)	-	-	-	-
Dognin et al. (2020)	VizWiz	25.88	21.49	49.17	72.45	16.23
Pasupuleti et al. (2021)	Flick8k	0.46(B-1) 0.24(B-2) 0.15(B-3) 0.06(B-4)	-	-	-	-
Ahsan et al. (2021)	VizWiz	22.3	-	45.0	53.8	14.1
Chharia and Upadhyay (2020)	Flick8k	0.86	-	-	-	-
Makav and Kılıç (2019a)	MSCOCO	57.9(B-1) 40.4(B-2) 27.9(B-3) 19.1(B-4)	-	-	60.0	-
Zaman et al. (2019)	Flick8k, MSCOCO	0.24(B-4)	-	-	-	-

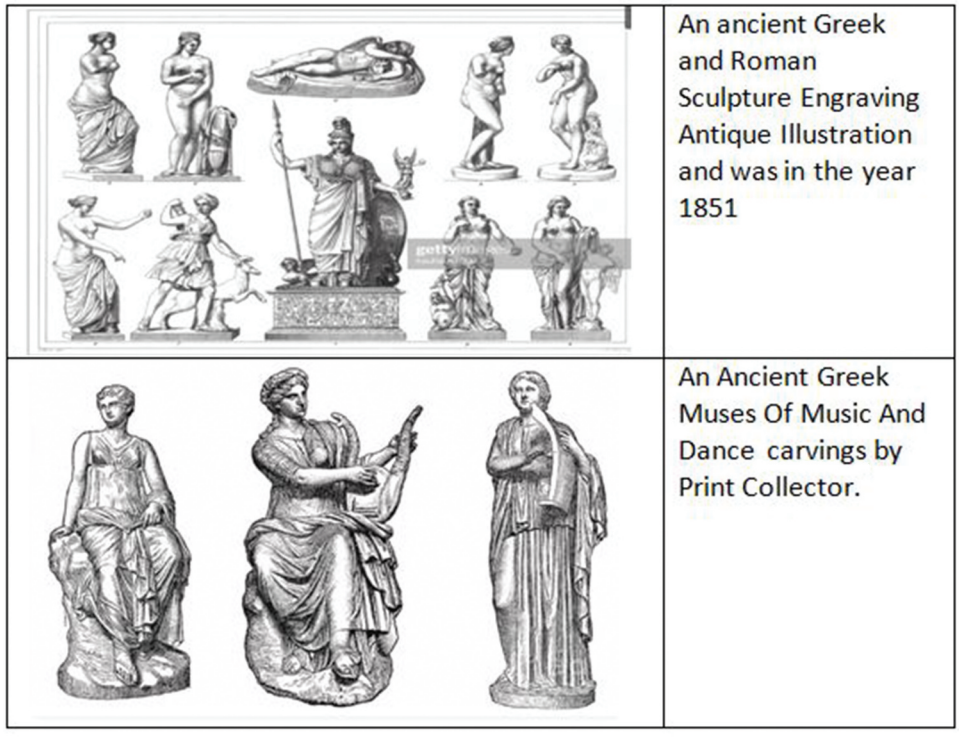


Figure 7. Image captioning with art images.

The performance of models for visually challenged persons is shown in Table 7.

We summarized the view of many researchers, to support a natural, socially important use case, i.e. presented their image captioning algorithms to generate captions from images which includes object detection, text detection and recognition. We outline the developed models that are

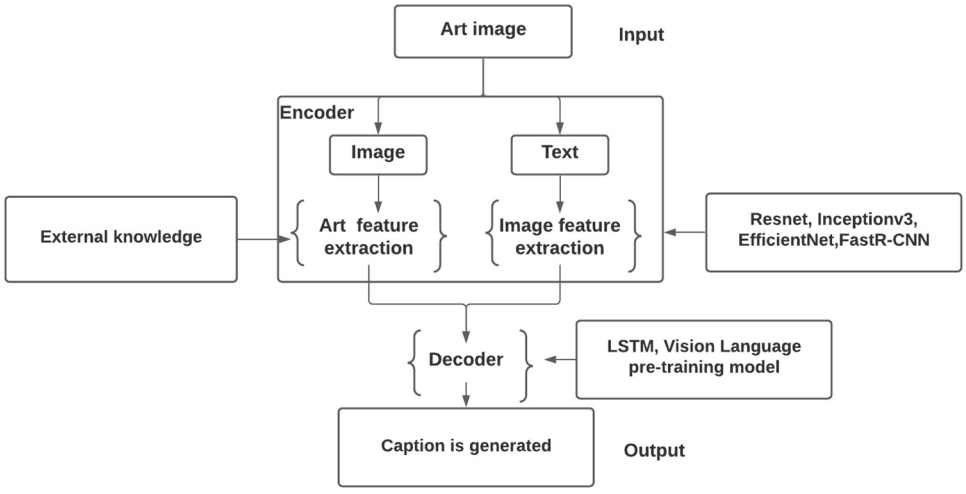


Figure 8. Workflow of image captioning for art images.

Table 8. Datasets used for art image captioning

Dataset	Number of samples	Description
IconClass Caption(Cetinic, 2021a)	87 K	The Icon Class Caption dataset is made up of photos taken from the Arkyves database, and it includes images of paintings, posters, sketches, prints, and manuscript pages.
SemArt(Garcia & Vogiatzis, 2018)	21 K	The SemArt dataset contains 21, 384 painting images, each of which is accompanied by an artistic comment as well as information such as artist, title, and date.
Ancient Egyptian art image captioning dataset(Sheng & Moens, 2019)	17 K	For the Ancient Egyptian art image captioning dataset, online sources include the Metropolitan, Brooklyn, and British Museums.
Ancient Chinese art image captioning dataset(Sheng & Moens, 2019)	7 K	The Metropolitan,Brooklyn and the British Museum have all contributed to the ancient Chinese art image captioning dataset.
WikiArt(Saleh & Elgammal, 2015)	85 K	WikiArt dataset consists of 85 K images but a subset of 52,562 images of paintings from the WikiArt collection are commonly used for captioning. Each image is tagged with a broad variety of labels like style, genre, artist, method, date of creation, and so on.
ArtWork(Cetinic, 2021b)	4 K	Artwork dataset consists of 4000 history based images across 9 iconographies like annunciation, adoration, baptism, still-life, nativity, virgin and child, rape, tower of babel and noli me tangere along with a description for each image.

Table 9. Analysis of various parameters used in art image captioning

Reference	Image extractor	Text extractor	Datasets	Evaluation metrics
Cetinic (2021a)	FasterR-CNN	vision-language pre-training model	Iconclass Caption	BLEU-1, BLEU-2 BLEU-3, BLEU-4, METEOR, ROUGE-L, CIDER.
Sheng and Moens (2019)	Resnet-18	LSTM	Ancient Egyptian art, Ancient Chinese art image captioning dataset	BLEU-1, BLEU-2 BLEU-3, BLEU-4, METEOR, ROUGE-L, CIDER,SPICE.
Cetinic (2021b)	Inceptionv3, EfficientNet	LSTM	Flickr8K, Flickr30K, ArtWork	BLEU-1, BLEU-2 BLEU-3, BLEU-4
Bai et al. (2021)	Faster R-CNN	vision-language pre-training model	Wikiart Dataset	BLEU-1, BLEU-2 BLEU-3, BLEU-4, METEOR, ROUGE-L, CIDER,CLIP-S, REFclip-S
Garcia and Vogiatzis (2018)	RESNET	LSTM	SemArt Dataset	BLEU-4,CIDER, METEOR, ROUGE-L, GreedyMatching, Skip-Thought, EmbeddingAverage

Table 10. Performance summarization in art image captioning. In this table GM, S-T, EA stands for GreedyMatching, Skip-Thought, EmbeddingAverage

Reference	Datasets	BLEU	METEOR	ROUGE	CIDER	SPICE	GM	S-T	EA
Cetinic (2021a)	Icon class Caption	14.8(BLEU-1) 12.8 (BLEU-2) 11.3 (BLEU-3) 10.0 (BLEU-4)	11.7	31.9	172.1	-	-	-	-
Sheng and Moens (2019)	ancient Egyptian art image captioning dataset	0.47(BLEU-1) 0.38 (BLEU-2) 0.33 (BLEU-3) 0.30 (BLEU-4)	0.20	0.44	1.87	0.29	-	-	-
	the ancient Chinese art image captioning daatset	0.54(BLEU-1) 0.45 (BLEU-2) 0.39 (BLEU-3) 0.35 (BLEU-4)	0.22	0.55	0.97	0.19	-	-	-
Cetinic (2021b)	ArtWork	22.47(BLEU-1) 12.02(BLEU-2) 7.95(BLEU-3) 6.27 (BLEU-4)	-	-	-	-	-	-	-
	Flickr8K	60.03(BLEU-1) 40.98(BLEU-2) 27.08(BLEU-3) 18.20(BLEU-4)	-	-	-	-	-	-	-
	Flickr30K	57.99(BLEU-1) 37.33(BLEU-2) 23.64(BLEU-3) 16.06(BLEU-4)	-	-	-	-	-	-	-
Bai et al. (2021)	Iconclass Caption	14.8(BLEU-1) 12.8 (BLEU-2) 11.3 (BLEU-3) 10.0 (BLEU-4)	11.7	31.9	172.1	-	-	-	-
Hacheme and Sayouti (2021)	SemArt Dataset	8.8(BLEU-4)	11.4	23.1	9.1	-	77.6	30.9	92.6

integrated into electronic devices which helps visually impaired people to traverse more easily. We review the work across various architectures and evaluated on different datasets.

3.3. Image captioning with art images

In the field of computer vision, generating captions from art images has been an important task (Figure 7). Artworks are characterized by various artistic styles, attributes, and motives along with great diversity of artists in different periods. To annotate historical artwork images, we require professional knowledge,high-level semantic information and the textual descriptions for ancient objects often include a lot of specific symbols. The most commonly used artwork datasets are IconClass Caption, SemArt, BibleVSA, ancient Egyptian art, the ancient Chinese, Flickr8K, Flickr30K, WikiArt dataset (Table 8).

An extensive quantitative and qualitative study (Bai et al., 2021; Garcia & Vogiatzis, 2018; Sheng & Moens, 2019) is used to validate the captions for the artwork. On these iconographies datasets, the authors in (Vaswani et al., 2017) fine-tuned the state-of-the-art models. External knowledge (Garcia & Vogiatzis, 2018) is utilised to characterise many characteristics of the image, such as its style, content, or composition (Table 9).

Art Image captioning workflow is depicted in Figure 8.

• An art image is used as the input for captioning.

• The encoder extracts the image characteristics using various CNN models. Art features needed to be extracted based on the external knowledge.

• Decoder combines both the art and image features using various models and generates the caption as output.

Art image captioning performance is summarized in Table 10. Generic image captioning metrics may not be highly correlated with the assessment of art expertise and originality, so the authors in (Hacheme & Sayouti, 2021) describe three other metrics and they are GreedyMatching, Skip-Thought and EmbeddingAverage.

We curated the work in image captioning for art images across various parameters. Annotating antique artwork photos requires professional skills. As artworks are characterized by various artistic styles, attributes, and motives. So, a comprehensive view of these models' performance w.r.t to various parameters helps to reduce the time to develop new applications.

3.4. Image captioning with fashion images

Fashion is an industry related to social, cultural, and economic implications in the real world. It is critical to provide correct captions for online fashion items not only to attract customers, but also to boost online sales. It is difficult to recognise and describe the rich features of fashion products, unlike conventional image captioning. In this case, the input is a fashion photograph, and the output is a fashion caption. The most commonly used datasets are DeepFashion, InFashAIv1, Fashion Captioning Dataset, FASHION-IQ, Fashion Database (Table 11). An instance of image captioning for fashion-related images is shown in Figure 9.

To increase the quality of text descriptions, Yang et al. (2020) proposes attribute-level and sentence-level semantic reward as measures, Hacheme and Sayouti (2021) combined DeepFashion and InFashAIv1 datasets for performance improvement. The authors in (Tateno et al., 2020) applied DNN to translate visual information collected from clothing into language expression, enaling visually impaired persons to access the shape and texture of objects. The authors in (J. Li et al., 2019) applied several hyperparameters to achieve a 39.12% average recall using a single model and a 43.67%a verage recall with an aggregation of 16 models on the FASHION-IQ dataset (Table 12).

Table 11. Datasets used for fashion image captioning

Dataset	Description
DeepFashion(Z. Liu et al., 2016)	The Deep Fashion dataset contains over 800,000 garment photos of tops and bottoms.
InFashAIv1(Hacheme & Sayouti, 2021)	InFashAI 15,716 image data is gathered from Pinterest and Afrikrea3 related to African fashion.Each image includes attributes like titles, prices, descriptions.
Fashion Captioning Dataset(Yang et al., 2020)	Fashion Captioning Dataset is composed with 993 K images. The images includes different age groups like kids, adults, old and different angles of images like front, back, top.
METEOR(Wu et al., 2021)	Fashion IQ dataset contains images related to dresses (19,087 images), shirts(31,728 images), and tops&tees(26,869 images) .Each image include textual descriptions, attribute labels like material, cost.
Fashion Database (Sadeh et al., 2019)	A Fashion Specialists team helped to acquire about 60 K outfit photographs for the Fashion Database. The assessment set consists of 500 photos with around 15 captions per image.

A detailed workflow of the process involved in fashion image captioning is shown in Figure 10 and the same is briefed below:

- The input for image captioning is a fashion image.
- Encoder extracts the features from the image using various CNN models.
- Decoder takes the input from encoder and also attributes of the image are taken into consideration to generate caption as output.

Figure 9. Image captioning with fashion images.

Table 12. Analysis of various parameters used in fashion image captioning

Reference	CNN	RNN	Datasets	Evaluation metrics
Hacheme and Sayouti (2021)	Resnet152	LSTM	InFashAIv1, DeepFashion	-
Wu et al. (2021)	EfficientNet-b7	Transformer	Fashion IQ	BLEU-4, ROUGE-L, CIDER, SPICE
Tateno et al. (2020)	VGG16	LSTM	DeepFashion	-
X. Li et al. (2021)	Resnet101	LSTM	DeepFashion	-
Yang et al. (2020)	Resnet101	LSTM	Fashion Captioning Dataset	BLEU-4, METEOR, ROUGE-L, CIDER, SPICE, mAP,AC
J. Li et al. (2019)	VGG	Transformer	FASHION-IQ	-
Sadeh et al. (2019)	Resnet	LSTM	Fashion Database	BLEU4, ROUGE-L, METEOR, CIDER-D, Div, Vocab.

Fashion image captioning performance is summarized in Table 13. Along with the generic image captioning metrics, the authors in (Yang et al., 2020) have used mAP (mean average precision) and ACC (accuracy), Sadeh et al. (2019) applied diversity (Div.), vocabulary usage (Vocab.)

Different from generic image captioning, identification and description of attributes plays a vital role in fashion image captioning. A detailed workflow of the process involved in fashion image captioning is also outlined. We briefed fashion image captioning architectures across various parameters. An intense view of the model performance w.r.t various parameters helps in generating accurate captions.

3.5. Image captioning with medical images

Medical captioning encodes medical images from a patient's examination and generates a full or partial report. Medical reports are always key decisive factors for initiating the right treatment of various diseases. Medical images are typically interpreted by highly skilled professionals. They write medical reports to describe the findings of the patient's abnormalities and diseases. Even for experienced radiologists, drafting a medical report can be time-consuming and unpleasant. As a result, the creation of medical reports can aid radiologists in making decisions, as well as assist medical teams in reducing workload and improve work efficiency. The most commonly used

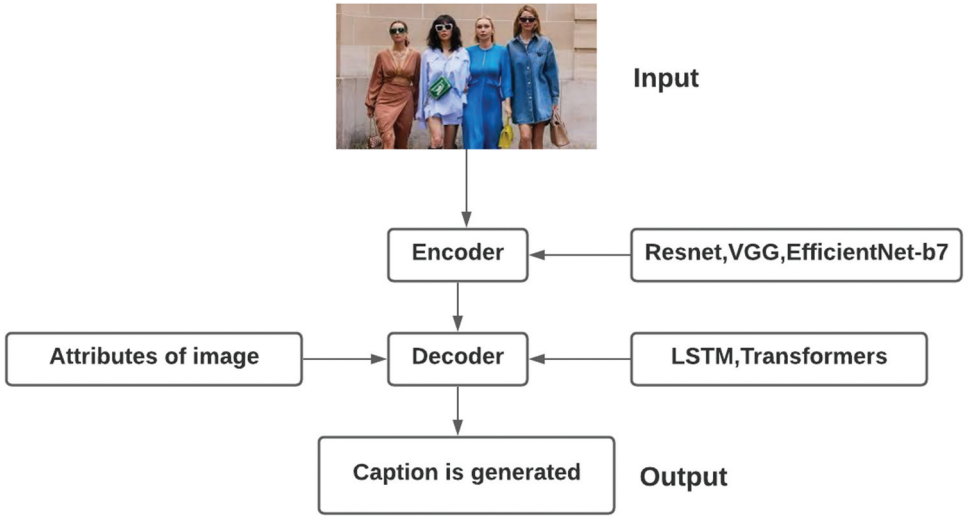


Figure 10. Workflow of image captioning for fashion images.

Table 13. Performance summarization in fashion image captioning. In this table mAP, ACC, Div, Vocab stands for mean average precision, accuracy, diversity, vocabulary usage

Reference	Datasets	BLEU-4	METEOR	ROUGE-L	CIDER	SPICE	mAP	ACC	Div	Vocab
Wu et al. (2021)	DeepFashion-Attr(dresses)	21.1	-	57.1	80.6	36.1	-	-	-	-
	DeepFashion-Attr(Shirts)	24.2	-	57.5	92.1	35.4	-	-	-	-
	DeepFashion–Attr(Tops & Tees) 22.1	-	55.4	82.3	35.0	-	-	-	-	-
Yang et al. (2020)	Fashion Captioning Dataset	24.2	23.2	47.2	114.8	21.9	0.240	0.512	-	-
Sadeh et al. (2019)	Fashion Database Model(GOOD)	0.341	0.55	0.29	0.438	-	-	-	0.965	0.281
	Fashion Database Model(TIP)	0.295	0.486	0.242	0.184	-	-	-	0.988	0.287

Table 14. Datasets used for medical image captioning

Dataset	Description
DeepEyeNet(Huang, Wu, Worring et al., 2021)	The DeepEyeNet dataset contains 1,811 Fluorescein Angiography photos in greyscale and 13,898 Color Fundus Photography images in colour. A clinical description is included with each retinal picture. The dataset is divided into three parts: 60% for training, 20% for validation,, and 20% for testing.
IU-X-RAY(Chen et al., 2020)	The Indiana University chest X-ray dataset, which contains 7470 chest X-ray scans and 3955 de-identified radiology reports.Each report includes Impression, Findings, and Indication.The dataset is divided into three parts: 70% for training, 10% validation and 20% for testing.
MIMIC-CXR(Chen et al., 2020)	MIMIC-CXR contains 377,110 chest X-ray images and 227,835 patient reports from a total of 64,588 patients. The training set has 368,960 records, the validation set has 2,991 records, and the test set has 5,159 records.
PEIR Gross(Jing et al., 2017)	The Pathology Education Informational Resource (PEIR) Gross dataset contains 7,442 images.
OPENI-IU	The OpenI-IU dataset contains 3,996 radiology reports and 8,121 accompanying chest X-ray images that have been carefully annotated with by human specialists. Only unique frontal photos and their related reports with findings or impressions are chosen from the dataset.
CheXpert (Yuan et al., 2019)	CheXpert has 224,316 multi-view chest x-ray images from 65,240 patients, representing 14 common radiographic findings.The observations are derived from radiology reports that are categorised as good, negative, or unsure using NLP methods.
OpenI(Wang et al., 2018)	OpenI is a radiography dataset of 3,851 distinct radiology reports and 7,784 frontal/lateral images. Body parts, observations, and diagnoses were annotated on each OpenI report.
CX-CHR(Jing et al., 2020)	A professional medical examination institution provided the CX-CHR dataset, which contains 35,500 pictures. Each image includes a textual report authored by trained radiologists, which includes parts including Complain, Findings, and Impression.
ChestX-ray14(Wang et al., 2018)	There are 108,948 frontal-view X-ray images in the ChestX-ray database. They divided the total dataset into 3 parts i.e. 70% to training, 10% to validation, and 10% to testing.

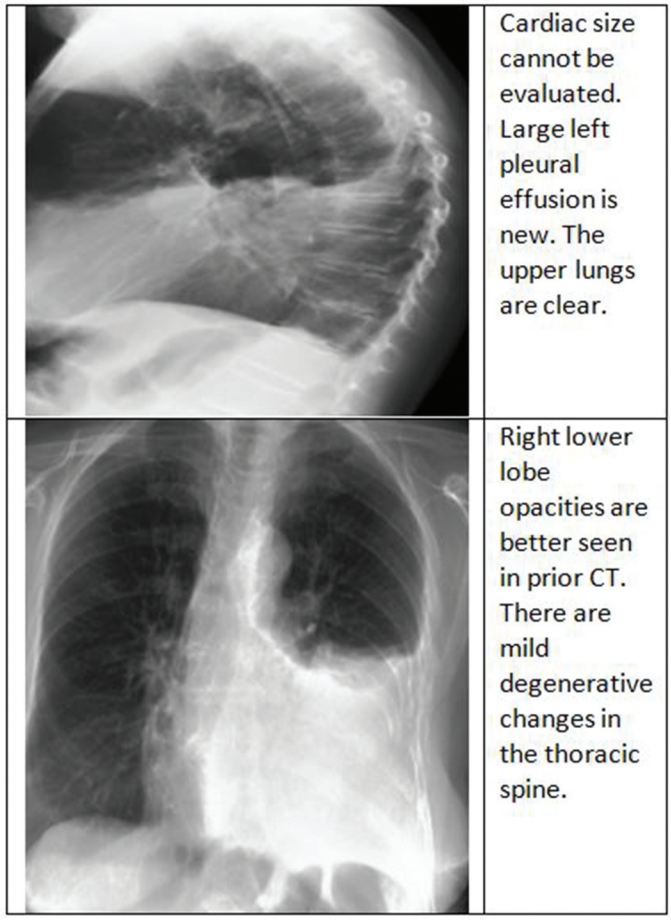


Figure 11. Image captioning with medical images.

datasets are DeepEyeNet, IU X-RAY, DAISI, CheXpert, Geneome, MIMIC-CXR, OpenI, CX-CHR, PeerGross, CXR (Table 14). An instance of medical captioning is shown in Figure 11.

With the tremendous amount of research publications in the medical domain, selection of relevant papers for review is a typical problem. We addressed this by prioritizing the very recent works based on the citation score (i.e. greater than 20).

In (Huang, Wu et al., 2021), the authors employed a multi-modal input encoder and a decoder architecture, Huang et al. (2021) introduced a retinal disease identifier (RDI) and a clinical description generator (CDG), Huang, Wu, Worring et al. (2021)applied contextualized keyword encoder and a medical description generator for retinal report generation (Table 15).

To generate radiology reports, for a given a set of radiology images (N), the visual backbone extracts the visual features F and results in the source sequence f1, f2, …, fs for the subsequent visual language model. The authors in(Chen et al., 2021, 2020; Jing et al., 2020; Liu, Ge et al., 2021; Liu, Wu et al., 2021; Liu, Yin et al., 2021; Pahwa et al., 2021; Wang et al., 2018, 2021; Xue et al., 2018; Yuan et al., 2019) extracted the visual features by pre-trained convolutional neural networks RESNET and also the authors in(Gale et al., 2018; Guanxiong Hsu Liu et al., 2019; Laserson et al., 2018; Y. Li et al., 2018; Nooralahzadeh et al., 2021; Pino et al., 2021; Zhou et al., 2021) used DenseNet as it is more effective in report generation task. The authour's in (Guanxiong Hsu Liu et al., 2019; Jing et al., 2017; Y. Li et al., 2018; Liu, Ge et al., 2021; Yuan et al., 2019) constructed a hierarchical LSTM model to generate the paragraph. As introduced by, (Chen et al., 2021, 2020; Ji et al., 2021; Nooralahzadeh et al., 2021; Pahwa et al., 2021) used transformers for the text

Table 15. Analysis of various parameters used in medical image captioning

Reference	CNN	RNN	Datasets	Evaluation metrics
Huang, Wu et al. (2021)	VGG16,VGG19, Resnet, Inceptionv3	Bidirectional LSTM	DeepEyeNet	ROUGE, BLEU:4, BLEU:3, BLEU:avg, CIDER, BLEU:2, BLEU:1.
Huang et al. (2021)	MobileNetV2, VGG16, VGG19, InceptionV3	LSTM	DeepEyeNet	BLEU:avg, BLEU:4, CIDER, BLEU:3, ROUGE, BLEU:2, BLEU:1.
Huang, Wu, Worring et al. (2021)	VGG16, VGG19	Transformers	DeepEyeNet	METEOR, BLEU:avg, BLEU:4, BLEU:3, CIDER, BLEU:2, BLEU:1, ROUGE:L .
Liu et al. (2021)	ResNet-50	Hierarchical LSTM	Indiana University chest X-ray,MIMIC: CXR	METEOR, BLEU:4, BLEU:3, ROUGE:L, BLEU:2, BLEU:1.
Wang et al. (2021)	Resnet-152	Hierarchical LSTM	Indiana University chest X-ray	METEOR BLEU:4 ROUGE BLEU:3 BLEU:2 CIDER BLEU:1.
Liu, Wu et al. (2021)	Resnet-152	LSTM	Indiana University chest X-ray,MIMIC: CXR	METEOR BLEU:4 BLEU:3 ROUGE:L BLEU:2 BLEU:1
Pahwa et al. (2021)	Resnet or VGGNET	Transformer	PEIR Gross, Indiana University chest X-ray	ROUGE BLEU:4 BLEU:3 METEOR BLEU:2 BLEU:1.
Zhou et al. (2021)	DenseNet-201	Bidirectional LSTM	Indiana University chest X-ray,MIMIC: CXR	CIDER BLEU:4 BLEU:3 ROUGE BLEU:2 BLEU:1 METEOR.
Pino et al. (2021)	Densenet-121	LSTM	Indiana University chest X-ray, MIMIC: CXR	CIDER BLEU:4, BLEU:3,ROUGE:L BLEU:2, BLEU:1.
Ji et al. (2021)	Joint Image Text Representation Learning Network	Transformer	MIMIC:CXR OPENI: IU	-
Chen et al. (2021)	ResNet101	Transformer	Indiana University chest X-ray	ROUGE:L BLEU:4, BLEU:3,METEOR, BLEU:2, BLEU:1 .
Nooralahzadeh et al. (2021)	DensNet-121	Transformer	Indiana University chest X-ray,MIMIC: CXR	METEOR BLEU1 BLEU2 ROUGE:L BLEU3 BLEU4 .
Liu, Ge et al. (2021)	ResNet-50	Hierarchical LSTM	Indiana University chest X-ray,MIMIC: CXR	ROUGE:L BLEU1 BLEU2 METEOR BLEU3 BLEU4
Yuan et al. (2019)	Resnet-152	Hierarchical LSTM	CheXpert	ROUGE BLEU:4, BLEU:3, METEOR BLEU:3, BLEU:4 .
Guanxiong Hsu Liu et al. (2019)	DenseNet	Hierarchical LSTM	MIMIC:CXR,OpenI	ROUGE:L, BLEU:4, BLEU:3, CIDER, BLEU:2, BLEU:1.
Y. Li et al. (2018)	DenseNet	Hierarchical LSTM	Indiana University chest X-ray,CX:CHR	ROUGE:L, BLEU:4, BLEU:3, BLEU:3, BLEU:4, CIDER.
Jing et al. (2017)	VGG19	Hierarchical LSTM	Indiana University chest X-ray, PeerGross	CIDER, ROUGE, BLEU:4, BLEU:3, BLEU:3, BLEU:4.

(*Continued*)

Reference	CNN	RNN	Datasets	Evaluation metrics
Xue et al. (2018)	Resnet-152	Bidirectional LSTM	Indiana University chest X-ray	ROUGE, METEOR, BLEU:4, BLEU:3, BLEU:3, BLEU:4
Laserson et al. (2018)	DenseNet	LSTM	Chest X:Ray	-
Gale et al. (2018)	DenseNet	LSTM	frontal pelvic X:rays	BLEU
Chen et al. (2020)	Resnet-101	Transformers	IUX:RAY, MIMIC:CXR	ROUGE:L METEOR, BLEU:4, BLEU:3
Jing et al. (2020)	ResNet-50	LSTM	IU:Xray,CX:CHR	BLEU:4 BLEU:3 BLEU:3 BLEU:4 ROUGE CIDER
Wang et al. (2018)	ResNet-50	LSTM	ChestX:ray14, OpenI	METEOR,BLEU:4, BLEU:3, ROUGE:L, BLEU:3, BLEU:4.

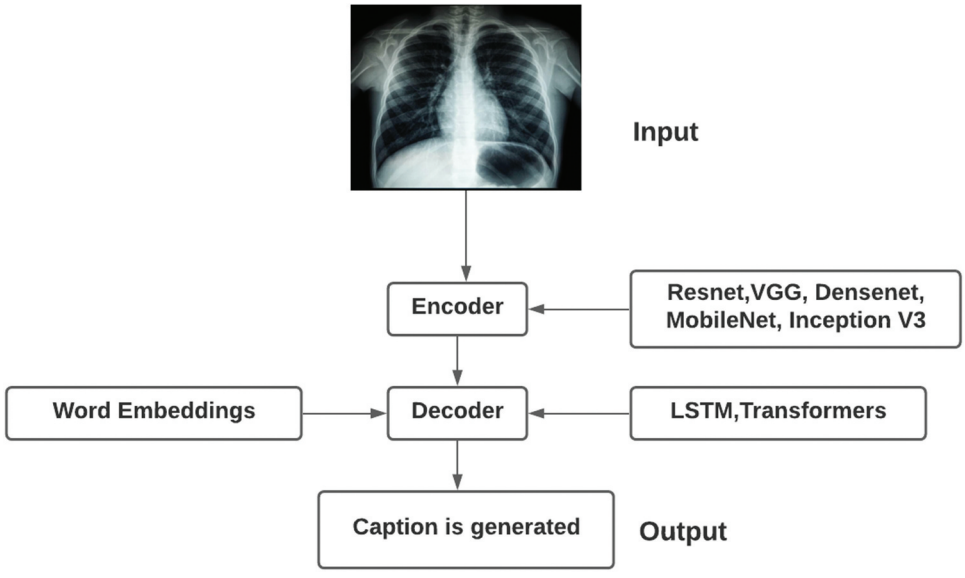


Figure 12. Workflow of image captioning for medical images.

generation of radiology reports. Xue et al. (2018) proposed a multimodal recurrent model containing an iterative decoder to improve the coherence between sentences. The authors Wang et al. (2018) proposed multi-attention model to combine image and text modalities using the CNN-RNN architecture, to improve disease classification and report generation, Jing et al. (2020) constructed a model to find the relationship between findings and impression.

The authors in(Liu, Yin et al., 2021) proposed contrastive learning to organize the data into similar/dissimilar image pairs. For chest X-ray report generation, the authors in (Guanxiong Hsu Liu et al., 2019; Y. Li et al., 2018; Liu, Wu et al., 2021) applied reinforcement learning and knowledge graph, Pahwa et al. (2021) used skip connections and transformers. The template-based multi-attention model (TMRGM) presented by Wang et al. (2021) for automatically creating reports for healthy and abnormal individuals. Most recently, Han et al. (2018); Zhang et al. (2020) utilized abnormality graph embedding module to assist the generation of reports. This is further extended by authors Gurari et al. (2020) to generate a report based on GAN's.

Table 16. Performance summarization in medical image captioning. In this table B:1, B:2, B:3, B:4 stands for BLEU:n

Reference	Datasets	BLEU	METEOR	ROUGE:L	CIDER
Huang, Wu et al. (2021)	DeepEyeNet	0.035(B:4) 0.074(B:3) 0.134 (B:2) 0.219(B:1)	-	0.252	0.398
Huang et al. (2021)	DeepEyeNet	0.032(B:4) 0.068(B:3) 0.114 (B:2) 0.184(B:1)	-	0.232	0.361
Huang, Wu, Worring et al. (2021)	DeepEyeNet	0.073(B:4) 0.100(B:3) 0.142 (B:2) 0.203(B:1)	0.188	0.211	0.389
Liu et al. (2021)	Indiana University chest X-ray	0.169(B:4) 0.222(B:3) 0.314 (B:2) 0.492(B:1)	0.193	0.381	-
	MIMIC:CXR	0.109(B:4) 0.152(B:3) 0.219 (B:2) 0.350(B:1)	0.151	0.283	-
Wang et al. (2021)	Indiana University chest X-ray	0.145(B:4) 0.201(B:3) 0.281 (B:2) 0.419(B:1)	0.183	0.280	0.359
Liu, Wu et al. (2021)	Indiana University chest X-ray	0.168(B:4) 0.224(B:3) 0.315 (B:2) 0.483(B:1)	-	0.376	0.351
	MIMIC:CXR	-	-	-	-
Pahwa et al. (2021)	PEIR Gross	0.148(B:4) 0.209(B:3) 0.278 (B:2) 0.399(B:1)	0.176	0.414	-
	Indiana University chest X-ray	0.467(B:1) 0.297(B:2) 0.214 (B:3) 0.162(B:4)	0.187	0.355	-
Zhou et al. (2021)	Indiana University chest X-ray	0.252(B:4) 0.314(B:3) 0.391 (B:2)0.536(B:1)	0.339	0.448	0.339
	MIMIC:CXR	0.372(B:1) 0.241(B:2) 0.168 (B:3) 0.123(B:4)	0.190	0.355	1.121
Pino et al. (2021)	Indiana University chest X-ray	0.273	-	0.352	0.249
	MIMIC:CXR	0.094	-	0.185	0.238
Chen et al. (2021)	Indiana University chest X-ray	0.170(B:4) 0.222(B:3) 0.309 (B:2) 0.475(B:1)	0.191	0.375	-
	MIMIC:CXR	0.106(B:4) 0.148(B:3) 0.218 (B:2) 0.353(B:1)	0.142	0.278	-
Nooralahzadeh et al. (2021)	Indiana University chest X-ray	0.173(B:4) 0.232(B:3) 0.317 (B:2) 0.486(B:1)	0.192	0.390	-
	MIMIC:CXR	0.107(B:4) 0.154(B:3) 0.232 (B:2) 0.378(B:1)	0.145	0.272	-

(*Continued*)

cogent engineering

Table16. (*Continued*)

Reference	Datasets	BLEU	METEOR	ROUGE:L	CIDER
Liu, Ge et al. (2021)	Indiana University chest X-ray	0.162(B:4) 0.217(B:3) 0.305 (B:2) 0.473(B:1)	0.186	0.378	-
	MIMIC:CXR	0.097(B:4) 0.140(B:3) 0.217 (B:2) 0.344(B:1)	0.133	0.281	-
Yuan et al. (2019)	CheXpert	0.278(B:4) 0.317(B:3) 0.380 (B:2) 0.500(B:1)	0.281	0.440	1.061
Guanxiong Hsu Liu et al. (2019)	MIMIC:CXR	0.104(B:4) 0.153(B:3) 0.223 (B:2) 0.352(B:1)	-	0.307	1.153
	OpenI	0.115(B:4) 0.171(B:3) 0.246 (B:2) 0.369(B:1)	-	0.359	1.490
Y. Li et al. (2018)	Indiana University chest X-ray	0.151(B:4) 0.208(B:3) 0.298 (B:2) 0.438(B:1)	-	0.322	0.343
	CX:CHR	0.486(B:4) 0.530(B:3) 0.587 (B:2) 0.673(B:1)	-	0.612	2.895
Jing et al. (2017)	Indiana University chest X-ray	0.247(B:4) 0.306(B:3) 0.386 (B:2) 0.517(B:1)	0.217	0.447	0.327
	PeerGross	0.113(B:4) 0.165(B:3) 0.218 (B:2) 0.300(B:1)	0.149	0.279	0.329
Xue et al. (2018)	Indiana University chest X-ray	0.195(B:4) 0.270(B:3) 0.358 (B:2) 0.464(B:1)	0.274	0.366	-
Chen et al. (2020)	Indiana University chest X-ray	0.165(B:4) 0.219(B:3) 0.304 (B:2) 0.470(B:1)	0.187	0.371	-
	MIMIC:CXR	0.103(B:4) 0.145(B:3) 0.218 (B:2) 0.353(B:1)	0.142	0.277	-
Jing et al. (2020)	Indiana University chest X-ray	0.166(B:4) 0.220(B:3) 0.290 (B:2) 0.401(B:1)	-	0.521	1.457
	CX:CHR	0.290(B:4) 0.323(B:3) 0.361 (B:2) 0.428(B:1)	-	0.504	2
Wang et al. (2018)	ChestX:ray14	0.0736(B:4) 0.1038(B:3) 0.1597(B:2) 0.2860(B:1)	0.1076	0.2263	-

cogent engineering

Table 17. Datasets used for other interdisciplinary domains

Dataset	Description
MS-COCO(Lin et al., 2014)	The MS-COCO database accommodate 123,287 images. There are five captions for each image.
LaRA(W. Li et al., 2020)	The dataset contains 11,178 images as traffic scenes with captions that can be used as a hint for prediction.
InstaPIC(Tan et al., 2019)	The InstaPIC dataset contains 648,761 training images and 5,000 testing images.
IAPR-ADD(Arriaga et al., 2017)	The authors created a dataset that contains 1008 captioned images. An image is an anomaly if it contains one of the following classes: knife, guns, fire, blood, dead bodies and broken objects.

A detailed workflow of the process involved in medical image captioning is shown in Figure 12 and the same is briefed below:

• The input for image captioning is a medical image.

• Encoder extracts the features from the image using various CNN models.

• Decoder takes the input from encoder and also attentive word embeddings are taken into consideration to generate the caption as output.

Medical captioning performance is summarized in Table 16. The automatic detection from medical images using ImageCLEF (Pelka et al., 2020) dataset achieved F1-scores of 0.3940 in 2020, F1-scores of 0.2823 in Imagemed Caption 2019, F1-scores of 0.1108 in ImageCLEFmed Caption 2018 and F1-scores of 0.1583 in ImageCLEFmed Caption 2017. The authors in (Allaouzi et al., 2018) presented an overview of datasets and metrics used for medical report generation, Pavlopoulos et al. (2021) illustrated different encoder decoder architectures, different evaluation measures and available datasets in the medical domain, Monshi et al. (2020) discussed the techniques involved in generating the radiology reports for the respective medical images.

We discussed about how to create medical reports more efficiently based on various parameters. Several architectures have been evaluated across different datasets. Substantial progress has been made towards implementing automatic reports based on various deep learning models. Medical reports help experienced radiologists to take a right decision for the treatment of the patients.

3.6. Image captioning in other domains

Image captioning is increasingly being employed in a variety of additional applications, including the effective retrieval of images in military applications, driving operations in complex traffic scenarios, and risky situations (Table 17).

The authors in (W. Li et al., 2020; Mori et al., 2019) utilized the model as an assistance system that can prevent traffic accidents, Arriaga et al. (2017) described how image captioning helps in the dangerous circumstances involving knife, guns, fire, blood, dead bodies and broken objects.

The summary work on other domains is shown in Table 18.

Table 18. Analysis of various parameters used in other interdisciplinary domains

Reference	Image extractor	Text extractor	Datasets	Evaluation metrics
Mori et al. (2019)	FastR-CNN	LSTM	MS-COCO	BLEU1, BLEU2, BLEU3, BLEU4, METEOR.
W. Li et al. (2020)	VGG-16	LSTM	LaRA	BLEU1, BLEU2, BLEU3, BLEU4.
Ghataoura and Ogbonnaya (2021)	InceptionV3	GRU	MS-COCO	-
Tan et al. (2019)	GoogLeNet	LSTM	MS-COCO InstaPIC	BLEU1, BLEU2, BLEU3, BLEU4, METEOR, SPICE, CIDER, ROUGE-L.
Arriaga et al. (2017)	Inception-V3	LSTM	IAPR-ADD (anomaly detection dataset)	BLEU,METEOR.

Table 19. Statistical analysis of various parameters used in image captioning models

Evaluation metric	Size		Architecture	
	p-VALUE	INTERPRETATION	p-VALUE	INTERPRETATION
BLEU-1	0.007071706	Dependent	0.000461338	Dependent.
BLEU-2	0.121710858	Independent	0.009153451	Dependent
BLEU-3	0.482669941	Independent	0.035984862	Dependent
BLEU-4	0.601058905	Independent	0.045192856	Dependent
METEOR	0.204939007	Independent	0.04129154	Dependent
ROUGE	0.000999074	Dependent	0.97849285	Independent
CIDER	0.328466992	Independent	1.14E-05	Dependent

3.7. Extracting insights from image captioning models across various domains

We compiled a list of image captioning models from various fields. However, a deeper knowledge of the model's performance in relation to numerous parameters remains unexplored. We use statistical analysis to solve this problem and reduce the time spent in brute forcing even for new applications.

Statistical analysis is a crucial tool in experimental research for effective interpretation. We have used Chi-square test for statistical analysis. Chi-Square test of independence is used to check whether two variables have a significant relationship between them. The Null hypothesis(H_0) and Alternate Hypothesis(H_1) are defined as below

H_0: The metric score is unaffected by size or architecture.

H_1:The metric score is influenced by size and architecture.

We tested the above hypothesis on our metadata which was created based on the different domains. From Table 19 the following conclusions can be drawn.

- **Interpretation 1**: The evaluation metrics BLEU-1, ROUGE are dependent on size.

• **Interpretation 2**: METEOR, BLEU-1, BLEU-2, CIDER, BLEU-3, and BLEU-4 are architecture-dependent evaluation metrics.

4. Challenges & research directions

4.1. Challenges

• Traditional image captioning models lack compositionality and naturalness since they frequently create captions in a sequential fashion, i.e., the next generated word is dependent on both the previous word and the image attribute. Even though it is syntactically correct, in some complex scenarios, semantically irrelevant language structures will be constructed.

• The second challenge is datasets, i.e., models struggle to differentiate captions across similar contexts when they overfit to some of the same objects that co-exist in a common domain.

• The third difficulty is determining the quality of the generated captions. Since the existing captioning models do not take the complete image context into account, the captions will not be helpful for the images that have high variance in comparison with training data.

4.2. Research directions

• Image captioning models have fewer datasets than other types of models. Furthermore, the annotation procedure is entirely manual. As a result, semi-automated or automatic annotation procedures are required.

• Image captioning models are evaluated using various evaluation metrics like BLEU, CIDER, ROUGE, etc. But the metrics vary from application to application. But there is no standard mechanism to opt for an appropriate image captioning metric for the application under consideration. This triggers the design of a framework for optimal metric selection considering various parameters like the dataset, application, and the model.

• Interdisciplinary image captioning models show very low performance and require more time. So there is a need for optimization techniques to improve performance.

• Image captioning models are data hungry. They require a lot of data to generate captions. There are recent deep learning advancements that can build robust models even with less data. Integration of these techniques in the context of image captioning would be a possible alternative for improving existing image captioning models.

• New approaches need to be developed to provide diverse, creative, and human-like captions across multiple areas.

5. Conclusion

Image captioning is now being applied across many domains. We observed that there is no single best architecture that performs best across many domains. To automatically guide the user to pick the right architecture for a new application, we surveyed, analysed, and interpreted various constraints associated with best-performed models across various domains. We summarised the aspects of datasets, architectures, and evaluation metrics, and also how the architectures are evaluated across different datasets.

At the end, we have given our inferences from the extensive survey, which would help the researchers to further streamline and reduce the burden of brute-forcing the highly complex neural network models. The interpretation can be further extended by considering multiple datasets across various domains.

Acknowledgements
We would like to thank VIT-AP University for facilitating the resources required for conducting this research.

Funding
The authors received no direct funding for this research.

Author details
Uddagiri Sirisha[1]
E-mail: sirisha.uddagiri@gmail.com
ORCID ID: http://orcid.org/0000-0003-2998-3402
Bolem Sai Chandana[1]
E-mail: saichandana.bolem@vitap.ac.in
[1] School of Computer Science & Engineering, VIT-AP University, Amaravathi, India.

Disclosure statement
No potential conflict of interest was reported by the authors.

Citation information
Cite this article as: Semantic interdisciplinary evaluation of image captioning models, Uddagiri Sirisha & Bolem Sai Chandana, *Cogent Engineering* (2022), 9: 2104333.

References

Adriyendi, A. (2021). A rapid review of image captioning. *Journal of Information Technology and Computer Science*, *6*(2), 158–169. https://doi.org/10.25126/jitecs.202162316

Ahsan, H., Bhalla, N., Bhatt, D., & Shah, K. (2021). Multi-modal image captioning for the visually impaired. arXiv preprint arXiv:2105.08106 53–60.

Allaouzi, I., Ben Ahmed, M., Benamrou, B., & Ouardouz, M. (2018). Automatic caption generation for medical images. *Proceedings of the 3rd International Conference on Smart City Applications* (pp. 1–6).

Anderson, P., Fernando, B., Johnson, M., & Gould, S. (2016). Spice: Semantic propositional image caption evaluation. *European conference on computer vision*. Springer (pp. 382–398).

Arriaga, O., Plöger, P., & Valdenegro-Toro, M. (2017). Image captioning and classification of dangerous situations. *arXiv preprint arXiv:1711.02578* https://doi.org/10.48550/arXiv.2105.08106.

Bai, S., & An, S. (2018). A survey on automatic image caption generation. *Neurocomputing*, *311*, 291–304. https://doi.org/10.1016/j.neucom.2018.05.080

Bai, Z., Nakashima, Y., & Garcia, N. (2021). Explain me the painting: Multi-topic knowledgeable art description generation. *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 5422–5432).

Banerjee, S., & Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization* (pp. 65–72).

Batra, V., He, Y., & Vogiatzis, G. (2018). Neural caption generation for news images. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (n.d.) A training algorithm for optimal margin classifiers. *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory* (pp. 144–152).

Bychkovsky, V., Paris, S., Chan, E., & Durand, F. (2011). Learning photographic global tonal adjustment with a database of input/output image pairs. *CVPR 2011*. IEEE (pp. 97–104).

Cetinic, E. (2021a). Iconographic image captioning for artworks. *International Conference on Pattern Recognition*. Springer. (pp. 502–516).

Cetinic, E. (2021b). Towards generating and evaluating iconographic image captions of artworks. *Journal of Imaging*, *7*(8), 123. https://doi.org/10.3390/jimaging7080123

Chen, C., Mu, S., Xiao, W., Ye, Z., Wu, L., & Ju, Q. (2019). Improving image captioning with conditional generative adversarial nets. *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, pp. 8142–8150).

Chen, J., & Zhuge, H. (2019). News image captioning based on text summarization using image as query. *2019 15th International Conference on Semantics, Knowledge and Grids (SKG)* (pp. 123–126). IEEE.

Chen, Z., song, Y., Chang, T.-H., & Wan, X. (2020). Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*.

Chen, Z., Shen, Y., Song, Y., & Wan, X. (2021). Cross- modal memory networks for radiology report generation. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 5904–5914).

Chharia, A., & Upadhyay, R. (2020). Deep recurrent architecture based scene description generator for visually impaired. *2020 12th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)* (pp. 136–141). IEEE.

Choi, S.-H., Jo, S. Y., & Jung, S. H. (2021). Component based comparative analysis of each module in image cap- tioning. *ICT Express*, *7*(1), 121–125. https://doi.org/10.1016/j.icte.2020.08.004

Chunseong Park, C., Kim, B., & Kim, G. (2017). Attend to you: Personalized image captioning with context sequence memory networks. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 895–903).

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. in '*2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)* (Vol. 1, pp. 886–893). Ieee.

Dognin, P., Melnyk, I., Mroueh, Y., Padhi, I., Rigotti, M., Ross, J., Schiff, Y., Young, R. A., & Belgodere, B. (2020), 'Image captioning as an assistive technology: Lessons learned from vizwiz 2020 challenge', *arXiv preprint arXiv:2012.11696*.

Elhagry, A., & Kadaoui, K. (2021). A thorough review on recent deep learning methodologies for image captioning. *arXiv preprint arXiv:2107.13114* (pp. 1–6).

Fang, H., Gupta, S., Iandola, F., Srivastava, R. K., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., and Platt, J. C. (2015). From captions to visual concepts and back. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1473–1482).

Furkan Biten, A., Gomez, L., Rusiñol, M., & Karatzas, D. (2019). Good news, everyone! context driven entity-aware captioning for news images. *arXiv e-prints* (pp. arXiv–1904).

Gale, W., Oakden-Rayner, L., Carneiro, G., Bradley, A. P., & Palmer, L. J. (2018). Producing radiologist-quality reports for interpretable artificial intelligence. *arXiv preprint arXiv:1806.00340*.

Gan, C., Gan, Z., He, X., Gao, J., & Deng, L. (2017). Stylenet: Generating attractive visual captions with styles. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3137–3146).

Garcia, N., & Vogiatzis, G. (2018). How to read paintings: Semantic art understanding with multi-modal

retrieval. *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*.

Ghataoura, D., & Ogbonnaya, S. (2021). Application of image captioning and retrieval to support military decision making. *2021 International Conference on Military Communication and Information Systems (ICMCIS)* (pp. 1–8). IEEE.

Grubinger, M., Clough, P., Müller, H., & Deselaers, T. (2006). The iapr tc-12 benchmark: A new evaluation resource for visual information systems. *International workshop ontoImage* (Vol. 2).

Gurari, D., Li, Q., Lin, C., Zhao, Y., Guo, A., Stangl, A., & Bigham, J. P. (2019). Vizwiz-priv: A dataset for recognizing the presence and purpose of private visual information in images taken by blind people. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 939–948).

Gurari, D., Zhao, Y., Zhang, M., & Bhattacharya, N. (2020). Captioning images taken by people who are blind. *European Conference on Computer Vision* (pp. 417–434). Springer.

Hacheme, G., & Sayouti, N. (2021). Neural fashion image captioning: Accounting for data diversity. *arXiv preprint arXiv:2106.12154*.

Han, Z., Wei, B., Leung, S., Chung, J., & Li, S. (2018). Towards automatic report generation in spine radiology using weakly supervised framework. *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 185–193). Springer.

Hodosh, M., Young, P., & Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, *47*, 853–899. https://doi.org/10.1613/jair.3994

Hossain, M. Z., Sohel, F., Shiratuddin, M. F., & Laga, H. (2019). A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (Csur)*, *51* (6), 1–36. https://doi.org/10.1145/3295748

Hu, A., Chen, S., & Jin, Q. (2020). Icecap: Information concentrated entity-aware image captioning. *Proceedings of the 28th ACM International Conference on Multimedia* (pp. 4217–4225).

Huang, J.-H., Wu, T.-W., & Worring, M. (2021). Contextualized keyword representations for multi-modal retinal image captioning. *Proceedings of the 2021 International Conference on Multimedia Retrieval* (pp. 645–652).

Huang, J.-H., Wu, T.-W., Yang, C.-H.-H., & Worring, M. (2021). Deep context-encoding network for retinal image captioning. *2021 IEEE International Conference on Image Processing (ICIP)* (pp. 3762–3766). IEEE.

Huang, J.-H., Yang, C.-H.-H., Liu, F., Tian, M., Liu, Y. C., Wu, T.-W., Lin, I., Wang, K., Morikawa, H., Chang, H. et al. (2021). Deepopht: Medical report generation for retinal images via deep models and visual explanation. *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 2442–2452).

Ji, Z., Shaikh, M. A., Moukheiber, D., Srihari, S. N., Peng, Y., & Gao, M. (2021). Improving joint learning of chest x-ray and radiology report by word region alignment. *International Workshop on Machine Learning in Medical Imaging* (pp. 110–119). Springer.

Jing, B., Xie, P., & Xing, E. (2017). On the automatic generation of medical imaging reports. *arXiv preprint arXiv:1711.08195*.

Jing, B., Wang, Z., & Xing, E. (2020). Show, describe and conclude: On exploiting the structure information of chest x-ray reports. *arXiv preprint arXiv:2004.12274*.

Johnson, J., Karpathy, A., & Fei-Fei, L. (2016). Densecap: Fully convolutional localization networks for dense captioning. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4565–4574).

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., Bernstein, M. S., & Fei-Fei, L. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, *123*(1), 32–73. https://doi.org/10.1007/s11263-016-0981-7

Laserson, J., Lantsman, C. D., Cohen-Sfady, M., Tamir, I., Goz, E., Brestel, C., Bar, S., Atar, M., & Elnekave, E. (2018). Textray: Mining clinical reports to gain a broad understanding of chest x-rays. *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 553–561). Springer.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324. https://doi.org/10.1109/5.726791

Li, Y., Liang, X., Hu, Z., & Xing, E. P. (2018). Hybrid retrieval-generation reinforced agent for medical image re- port generation Proceedings of the 32nd International Conference on Neural Information Processing Systems*31*, 1537-–1547 https://doi.org/10.48550/arXiv.1805.08298.

Li, J., Lee, J.-W., Song, W.-S., Shin, K.-Y., & Go, B.-H. (2019). Designovel's system description for fashion-iq challenge 2019. *arXiv preprint arXiv:1910.11119* https://doi.org/10.48550/arXiv.1910.11119.

Li, W., Qu, Z., Song, H., Wang, P., & Xue, B. (2020). The traffic scene understanding and prediction based on image captioning. *IEEE Access*, *9* , 1420–1427. https://doi.org/10.1109/ACCESS.2020.3047091

Li, X., Ye, Z., Zhang, Z., & Zhao, M. (2021). Clothes image caption generation with attribute detection and visual attention model. *Pattern Recognition Letters*, *141*, 68–74. https://doi.org/10.1016/j.patrec.2020.12.001

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. *Text summarization branches out* (pp. 74–81).

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014) Microsoft coco: Common objects in context. *European conference on computer vision* (pp. 740–755). Springer.

Liu, Z., Luo, P., Qiu, S., Wang, X., & Tang, X. (2016). Deepfashion: Powering robust clothes recognition and re- trieval with rich annotations. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1096–1104).

Liu, G. H., Boag, T.-M. H.-M.-M., Weng, W., Szolovits, W.-H., & Ghassemi, M. (2019). Clinically accurate chest x-ray report generation. *Machine learning for healthcare conference* (pp. 249–269). PMLR.

Liu, X., Xu, Q., & Wang, N. (2019). A survey on deep neural network-based image captioning. *The Visual Computer*, *35*(3), 445–470. https://doi.org/10.1007/s00371-018-1566-y

Liu, F., Wang, Y., Wang, T., & Ordonez, V. (2020). Visual news: Benchmark and challenges in news image captioning. *arXiv preprint arXiv:2010.03743* https://doi.org/10.48550/arXiv.2010.03743.

Liu, F., Ge, S., & Wu, X. (2021). Competence-based multimodal curriculum learning for medical report generation. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural*

Language Processing (Volume 1: Long Papers) (pp. 3001–3012).

Liu, F., Wu, X., Ge, S., Fan, W., & Zou, Y. (2021). Exploring and distilling posterior and prior knowledge for radiology report generation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 13753–13762).

Liu, F., Yin, C., Wu, X., Ge, S., Zhang, P., & Sun, X. (2021). Contrastive attention for automatic chest x-ray report generation. *arXiv preprint arXiv:2106.06965* https://doi.org/10.48550/arXiv.2106.06965.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94

Makav, B., & Kılıç, V. (2019a). A new image captioning approach for visually impaired people. *2019 11th International Conference on Electrical and Electronics Engineering (ELECO)* (pp. 945–949). IEEE.

Makav, B., & Kılıç, V. (2019b). Smartphone-based image captioning for visually and hearing impaired. *2019 11th International Conference on Electrical and Electronics Engineering (ELECO)* (pp. 950–953). IEEE.

Monshi, M. M. A., Poon, J., & Chung, V. (2020). Deep learning in generating radiology reports: A survey. *Artificial Intelligence in Medicine*, *106*, 101878. https://doi.org/10.1016/j.artmed.2020.101878

Mori, Y., Fukui, H., Hirakawa, T., Nishiyama, J., Yamashita, T., & Fujiyoshi, H. (2019). Attention neural baby talk: Captioning of risk factors while driving. *2019 IEEE Intelligent Transportation Systems Conference (ITSC)* (pp. 4317–4322). IEEE.

Nooralahzadeh, F., Gonzalez, N. P., Frauenfelder, T., Fujimoto, K., & Krauthammer, M. (2021). Progressive transformer- based generation of radiology reports. *arXiv preprint arXiv:2102.09777* https://doi.org/10.48550/arXiv.2102.09777.

Oluwasammi, A., Aftab, M. U., Qin, Z., Ngo, S. T., Doan, T. V., Nguyen, S. B., Nguyen, S. H., Nguyen, G. H., & Selisteanu, D. (2021). Features to text: A comprehensive survey of deep learning on semantic segmentation and image captioning. *Complexity*, *2021*, 1–19. https://doi.org/10.1155/2021/5538927

Pahwa, E., Mehta, D., Kapadia, S., Jain, D., & Luthra, A. (2021). Medskip: Medical report generation using skip connections and integrated attention. *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 3409–3415).

Pan, J.-Y., Yang, H.-J., Faloutsos, C., & Duygulu, P. (2004). Gcap: Graph-based automatic image captioning. *2004 Conference on Computer Vision and Pattern Recognition Workshop*. IEEE.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311–318).

Pasupuleti, S., Dadi, L., Gadi, M., & Krishnaveni, R. (2021). Image recognition and voice translation for visually impaired. *International Journal of Research in Engineering, Science and Management*, *4*(5), 18–23 http://www.journals.resaim.com/ijresm/article/view/713.

Pavlopoulos, J., Kougia, V., Androutsopoulos, I., & Papamichail, D. (2021). Diagnostic captioning: A survey. *arXiv preprint arXiv:2101.07299* 1–32 https://doi.org/10.48550/arXiv.2101.07299 .

Pelka, O., Friedrich, C. M., García Seco de Herrera, A., & Müller, H. (2020). Overview of the imageclefmed 2020 concept prediction task. *Proceedings of the CLEF 2020- Conference and labs of the evaluation forum*. number CONFERENCE, 22-25 September 2020.

Pino, P., Parra, D., Besa, C., & Lagos, C. (2021). Clinically correct report generation from chest x-rays using templates. *International Workshop on Machine Learning in Medical Imaging* pp. 654–663. Springer.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., & Lazebnik, S. (2015). Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *Proceedings of the IEEE international conference on computer vision* (pp. 2641–2649).

Ramisa, A., Yan, F., Moreno-Noguer, F., & Mikolajczyk, K. (2017). Breakingnews: Article annotation by image and text processing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(5), 1072–1085. https://doi.org/10.1109/TPAMI.2017.2721945

Rane, C., Lashkare, A., Karande, A., & Rao, Y. (2021). Image captioning based smart navigation system for visually impaired. *2021 International Conference on Communication information and Computing Technology (ICCICT)* (pp. 1–5). IEEE

Sadeh, G., Fritz, L., Shalev, G., & Oks, E. (2019). Generating diverse and informative natural language fashion feedback. *arXiv preprint arXiv:1906.06619* https://doi.org/10.48550/arXiv.1906.06619.

Saleh, B., & Elgammal, A. (2015). Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *arXiv preprint arXiv:1505.00855* https://doi.org/10.48550/arXiv.1505.00855.

Sheng, S., & Moens, M.-F. (2019). Generating captions for images of ancient artworks. *Proceedings of the 27th ACM International Conference on Multimedia* (pp. 2478–2486).

Staniūtė, R., & Šešok, D. (2019). A systematic literature review on image captioning. *Applied Sciences*, *9*(10), 2024. https://doi.org/10.3390/app9102024

Stefanini, M., Cornia, M., Baraldi, L., Cascianelli, S., Fiameni, G., & Cucchiara, R. (n.d.) From show to tell: A survey on image captioning. *arXiv preprint arXiv:2107.06912* pp. 1–27. https://doi.org/10.48550/arXiv.2107.06912.

Tan, J. H., Chan, C. S., & Chuah, J. H. (2019). Comic: Toward a compact image captioning model with attention. *IEEE Transactions on Multimedia*, *21*(10), 2686–2696. https://doi.org/10.1109/TMM.2019.2904878

Tateno, K., Takagi, N., Sawai, K., Masuta, H., & Motoyoshi, T. (2020) Method for generating captions for clothing images to support visually impaired people. *2020 Joint 11th International Conference on Soft Computing and Intelligent Systems and 21st International Symposium on Advanced Intelligent Systems (SCIS-ISIS)* (pp. 1–5). IEEE.

Tran, K., He, X., Zhang, L., Sun, J., Carapcea, C., Thrasher, C., Buehler, C., & Sienkiewicz, C. (2016). Rich image captioning in the wild. *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 49–56).

Tran, A., Mathews, A., & Xie, L. (2020). Transform and tell: Entity-aware news image captioning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 13035–13045).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30* https://doi.org/10.48550/arXiv.1706.03762.

Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). Cider: Consensus-based image description evaluation. *Proceedings of the IEEE conference on*

computer vision and pattern recognition (pp. 4566–4575).

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3156–3164).

Wang, X., Peng, Y., Lu, L., Lu, Z., & Summers, R. M. (2018). Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 9049–9058).

Wang, Y., Xu, J., Sun, Y., & He, B. (2019). Image captioning based on deep learning methods: A survey. *arXiv preprint arXiv:1905.08110* pp. 1–7 https://doi.org/10.48550/arXiv.1905.08110.

Wang, X., Zhang, Y., Guo, Z., & Li, J. (2021). Tmrgm: A template-based multi-attention model for x-ray imaging report generation. *Journal of Artificial Intelligence for Medical Sciences*, *2*(1–2), 21–32. https://doi.org/10.2991/jaims.d.210428.002

Wu, H., Gao, Y., Guo, X., Al-Halah, Z., Rennie, S., Grauman, K., & Feris, R. (2021). Fashion iq: A new dataset towards retrieving images by natural language feedback. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 11307–11317).

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. *International conference on machine learning* pp. 2048–2057. PMLR.

Xue, Y., Xu, T., Rodney Long, L., Xue, Z., Antani, S., Thoma, G. R., & Huang, X. (2018). Multimodal recurrent model with attention for automated radiology report generation. *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 457–466). Springer,

Yang, X., Zhang, H., Jin, D., Liu, Y., Wu, C.-H., Tan, J., Xie, D., Wang, J., & Wang, X. (2020). Fashion captioning: Towards generating accurate descriptions with semantic rewards. *European Conference on Computer Vision* (pp. 1–17). Springer.

Yang, Z., & Okazaki, N. (2020). Image caption generation for news articles. *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 1941–1951).

Yang, X., Karaman, S., Tetreault, J., & Jaimes, A. (2021). Journalistic guidelines aware news image captioning. *arXiv preprint arXiv:2109.02865* https://doi.org/10.48550/arXiv.2109.02865.

You, Q., Jin, H., Wang, Z., Fang, C., & Luo, J. (2016). Image captioning with semantic attention. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4651–4659).

Yuan, J., Liao, H., Luo, R., & Luo, J. (2019). Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. *International Conference on Medical Image Computing and Computer- Assisted Intervention* (pp. 721–729). Springer.

Zaman, S., Abrar, M. A., Hassan, M. M., & Islam, A. N. (2019). A recurrent neural network approach to image captioning in braille for blind-deaf people. *2019 IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON)* (pp. 49–53). IEEE.

Zhang, Y., Wang, X., Xu, Z., Yu, Q., Yuille, A., & Xu, D. (2020). When radiology report generation meets knowledge graph. *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, pp. 12910–12917).

Zhao, W., Hu, Y., Wang, H., Wu, X., & Luo, J. (2021). Boosting entity-aware image captioning with multi-modal knowledge graph. *arXiv preprint arXiv:2107.11970*. https://doi.org/10.48550/arXiv.2107.11970.

Zhou, Y., Huang, L., Zhou, T., Fu, H., & Shao, L. (2021). Visual-textual attentive semantic consistency for medical report generation. *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 3985–3994).

Cogent Engineering **(ISSN: 2331-1916) is published by Cogent OA, part of Taylor & Francis Group.**