Artificial Intelligence Review (2023) 56:13619–13661
https://doi.org/10.1007/s10462-023-10488-2

A comprehensive survey on image captioning: from handcrafted to deep learning-based techniques, a taxonomy and open research issues

Himanshu Sharma[1] · **Devanand Padha**[1]

Published online: 17 April 2023

Abstract
Image captioning is a pretty modern area of the convergence of computer vision and natural language processing and is widely used in a range of applications such as multi-modal search, robotics, security, remote sensing, medical, and visual aid. The image captioning techniques have witnessed a paradigm shift from classical machine-learning-based approaches to the most contemporary deep learning-based techniques. We present an in-depth investigation of image captioning methodologies in this survey using our proposed taxonomy. Furthermore, the study investigates several eras of image captioning advancements, including template-based, retrieval-based, and encoder-decoder-based models. We also explore captioning in languages other than English. A thorough investigation of benchmark image captioning datasets and assessment measures is also discussed. The effectiveness of real-time image captioning is a severe barrier that prevents its use in sensitive applications such as visual aid, security, and medicine. Another observation from our research is the scarcity of personalized domain datasets that limits its adoption into more advanced issues. Despite influential contributions from several academics, further efforts are required to construct substantially robust and reliable image captioning models.

Keywords Attention-based image captioning · Encoder-decoder architecture · Image captioning · Multimodal embedding

1 Introduction

One of the fundamental abilities of humans is the potential to detect and comprehend a succinct description of the prominent components of an image using natural language. With such powerful abilities, it is incredibly simple to predict, and correlate a description with every image we encounter. However, making machines imitate such human expertise with reliable precision level is the most challenging research problem. A sentence that concisely describes the contents of an image is called the image caption as shown in Fig. 1.

✉ Himanshu Sharma
himanshusharma.csit@gmail.com

[1] Department of Computer Science and Information Technology, Central University of Jammu, Jammu, Jammu & Kashmir 181124, India

Fig. 1 An image with a possible set of captions from the Flickr 8K (Hodosh et al. 2013) dataset

The Image Captioning (IC) framework is thus an artificially intelligent model that generates captions for any given query image. The generated captions are short enough to fit into a single sentence and long enough to describe all the salient contents of the image. IC is considered as an extension to conventional image classification and image annotation systems that associate a single or multiple labels with the given image. IC models are usually more detailed than these conventional systems in terms of describing the image. An IC model not only classifies images based on their visual contents but also describes their semantic details, as well as the attributes and spatial relationships that exist between visual elements. The type of captions generated by these IC models can also vary depending on the domains for which they are trained. A road navigation captioning system (Li et al. 2021) for instance explains the road structures, traffic conditions, and signal status. Similarly, a content retrieval system (Verma and Jawahar 2014; Wang et al. 2020) retrieves the multi-modal web contents as requested by the users in the search query. The widespread implications of deep learning in critical domains such as medical image analysis (Bhosale et al. 2022), news captioning (Yumeng et al. 2021), and portable health monitoring systems (Bhosale and Patnaik 2022) raise the demand for IC frameworks further, as IC frameworks can sort and characterize images based on multimodal queries. In the present era, when an exponential amount of medical imaging is being produced every day (Bhosale and Patnaik 2022), IC can be especially helpful in saving a lot of human effort and time by interpreting, sorting, categorizing, and classifying crucial images automatically.

The visual detector and the description generation model are two architectural sub-components of an IC model. A visual detection sub-component recognizes and identifies the visual contents of the query image and generates an encoded representation for the inferred visual details known as context vectors. The description generation sub-component follows the visual detector and decodes the context vectors by describing the visual details contained in them using natural language. The visual detector often consists of object detection models such as a Convolutional Neural Network (CNN) or global and local feature extractors such as Scale-Invariant Feature Transform (SIFT), Global Image Descriptor (GIST), and Histogram of Oriented Gradients (HOG). The description generation model decomposes the input context into the captions using sequential models such as Recurrent Neural Network (RNN) and Long Short Term Memory (LSTM). Both the sub-components perform significant service in IC and hence the accuracy of an IC model depends directly on the performance of both the individual models. Earlier, machine learning-based approaches were used for object detection and caption generation. Later, deep learning-based models such as CNN, RNN, and LSTMs replaced the classical techniques. IC thus combines two state-of-the-art fields of Artificial Intelligence (AI) namely Computer Vision (CV) and Natural Language Processing (NLP). The domain of IC hence lies in the application areas of both CV and NLP.

IC is used for a wide range of applications, including vision aid systems, multi-modal content retrieval systems, robotic vision, remote sensing, and medical imaging. IC has expanded fast in the previous ten years, with deep learning techniques supporting its progress. There has been a significant deal of research on this issue, and as a consequence, many research articles with a wide range of experimental combinations in terms of CV and NLP methodologies have been published. There have also been attempts to produce non-English captions, as well as captions tailored to certain domains such as remote sensing, medical, social media, and journalism. With the significant rise of IC research, it is critically necessary to comprehensively survey IC literature.

Despite the widespread interest in the topic, only a few survey studies (Bernardi et al. 2016; Kumar and Goel 2018; Liu et al. 2019; Bai and An 2018; Hossain et al. 2018; Li et al. 2019; Alam et al. 2020; Amirian et al. 2020; Zohourianshahzadi and Kalita 2021) have been published. Although prior studies provided a thorough and comparative assessment of IC approaches, these surveys could only cover a subset of IC techniques because most state-of-the-art models were released after these surveys. As indicated in Table 1, no current research has analyzed the innovative application areas of IC such as non-English IC, remote sensing IC, and medical IC jointly. As a result, we propose to fill the

Table 1 Comparison of existing surveys with our survey

Authors	Models surveyed						
	Machine learning-based		Deep-learning-based				
	Generation approaches	Retrieval approaches	Encoder-decoder	Attention-based approaches	Non-English captioning	Remote sensing captioning	Medical captioning
Bernardi et al. (2016)	✓	✓	✓	✓	✗	✗	✗
Kumar and Goel (2018)	✓	✓	✓	✓	✗	✗	✗
Liu et al. (2019)	✓	✗	✓	✓	✗	✗	✗
Bai and An (2018)	✓	✓	✓	✓	✗	✗	✗
Hossain et al. (2018)	✗	✗	✓	✓	✗	✗	✗
Li et al. (2019)	✗	✗	✓	✓	✗	✗	✗
Alam et al. (2020)	✗	✗	✓	✓	✗	✗	✗
Amirian et al. (2020)	✗	✗	✓	✓	✗	✗	✗
Zohourianshahzadi and Kalita (2021)	✗	✗	✗	✓	✗	✗	✗
Our survey	✓	✓	✓	✓	✓	✓	✓

aforementioned research gaps by undertaking a comprehensive review of IC, encompassing all generations and methodologies in a single study. The abbreviations used throughout this survey study are listed below in Table 2.

To begin with, the scope, coverage, and limitations of this survey are discussed in Sect. 2. In Sect. 3, we present two alternative IC taxonomies categorizing all IC models based on their architecture and applications. We go over the classical machine learning-based and deep learning-based approaches of IC in Sects. 4 and 5 respectively. In Sect. 6, we discuss IC models based on their applicability. Sections 7 and 8 provide an overview of benchmark datasets and evaluation metrics of IC. Finally, in Sects. 9 and 10, the open research challenges and conclusions are discussed.

2 Scope and coverage

Our study comprises scholarly articles published by ACM, ACL, Elsevier, IEEE, MDPI, Springer, and others in well-known transactions, journals, and conference proceedings. Table 10 of the appendix contains extensive information about the kinds of articles and their publishers cited in this survey. Additionally, Tables 11 and 12 in the appendix indicate the frequency distribution of different publishers and article types mentioned in this research. We investigate the origins of IC and its numerous implications in a wide variety of potential areas. Figure 2a shows the frequency distribution of model architectures studied in the current and past surveys, and Fig. 2b shows the distribution of article publishers cited in our survey. Handcrafted feature engineering based on machine learning techniques was used in early IC research. The adoption of deep learning-based models such as CNN, RNN, and LSTMs transformed the paradigm in 2013.

The scope of this review has however been limited in certain directions due to the rapid expansion of the IC literature driven by ongoing research on deep learning. This survey's major objective is to investigate and cover all core architectural generations of IC frameworks. Since ensembled and compositional architectures of IC frameworks comprise merely a mixture of these core architectures, we confine our research to them. In addition, instead of delving comprehensively into the different image captioning types like dense captioning, stylized captioning, and scene captioning, our research concentrates more on the operational and functional principles of IC frameworks. Consequently, future survey studies may be directed on these limitations to expand the IC literature beyond this study's scope.

3 Taxonomy of image captioning

There has been a considerable amount of effort done towards classifying different IC models. Many differentiating features, such as model architectures, the kind of visual detection and caption generating sub-component, the type of captions being produced, and the learning employed to train IC models guide these classifications. Bernardi et al. (2016) proposed one of the first IC taxonomies, categorizing IC models into two types namely generation models and retrieval-based models. Kumar and Goel (2018) classified IC models based on the methodologies (machine learning and deep learning) used in their creation. The existing IC models in their survey are classified into two types namely template-based and neural network-based. Liu et al. (2019) categorized current literature into three

Table 2 A list of abbreviations and their meanings used in this survey

S. no.	Abbreviation	Meaning	S. no.	Abbreviation	Meaning
1	AI	Artificial Intelligence	27	MELM	Maximum Entropy Language Model
2	AIC	Attention-based Image Captioning	28	METEOR	Metric for Evaluation of Translation with Explicit Ordering
3	b-IDF	binary-Inverse Document Frequency	29	MIC	Medical Image Captioning
4	BLEU	Bilingual Evaluation Understudy	30	MIL	Multiple Instances Learning
5	BRNN	Bidirectional Recurrent Neural Network	31	MLBL-B	Modality-based Log Bilinear Model
6	CAA	Canonical Correlation Analysis	32	MLBL-F	Factored 3-way Log Bilinear Model
7	CGAN	Conditional Generative Adversarial Networks	33	MRNN	Multimodal Recurrent Neural Network
8	CIDEr	Consensus-based Image Description Evaluation	34	MS-COCO	Microsoft-Common Objects in Context
9	CNN	Convolutional Neural Network	35	NLP	Natural Language Processing
10	CoSMoS	Common Subspace of Models and Similarity	36	RBF	Radial Basis Function
11	CRF	Conditional Random Field	37	RCNN	Regional Convolutional Neural Network
12	CV	Computer Vision	38	RNN	Recurrent Neural Network
13	DT-RNN	Dependency Tree- Recurrent Neural Network	39	ROUGE	Recall Oriented Understudy for Gisting Evaluation
14	GAN	Generative Adversarial Network	40	RSIC	Remote Sensing Image Captioning
15	GCN	Graph Convolutional Network	41	RL	Reinforcement Learning
16	GIC	Generic purpose Image Captioning	42	SAE	Stack-Auxiliary Embedding
17	GIST	Global Image Descriptor	43	SC-NLM	Structured Context Vector Neural Language Model
18	GNN	Graph Neural Network	44	SEM	Scanning Electron Microscope
19	GRU	Gated Recurrent Unit	45	SIFT	Scale-Invariant Feature Transform
20	HMM	Hidden Markov Model	46	SMT	Statistical Mechanical Translation
21	HOG	Histogram of Oriented Gradients	47	SPICE	Semantic Proportional Image Captioning Evaluation
22	IC	Image Captioning	48	SVM	Support Vector Machine
23	ILP	Integer Linear Programming	49	UAV	Unmanned Aerial Vehicle
24	KCAA	Kernel Canonical Correlation Analysis	50	VDG	Visual Dependency Grammar
25	LCRN	Long-term Recurrent Convolutional Network	51	VDR	Visual Dependency Representation
26	LSTM	Long Short Term Memory	52	DLCT	Dual- Level CollaborativeTransformer

categories: template-based, retrieval-based, and end-to-end based. Bai and An (2018) grouped IC models into early works (template-based and retrieval-based models) and neural network-based models. Hossain et al. (2018) characterize one of the concise and extensive taxonomies of deep learning-based IC models, organizing them into six different types based on their architecture, the visual sub-component, the language sub-component, the type of captions generated, and the learning used to train the models. IC is such a vast scientific domain and hence there are a variety of ways of categorizing the existing literature. The two most significant classifications are based on model architectures and application areas. In this survey, we suggest an alternate IC taxonomy based on architecture (shown in Fig. 3) and a novel application-specific IC taxonomy (shown in Fig. 4).

At the most basic level, IC models may be divided into two types based on whether they employ handcrafted machine learning techniques such as HOG, SIFT, GIST, and Support Vector Machine (SVM) or deep learning-based approaches such as CNN and LSTM. Traditional machine-learning-based IC techniques are further classified into two types namely generation-based approaches and retrieval-based approaches. The generation methodologies further consist of statistical language models and template-based techniques. The former uses statistical language models such as n-grams for generating captions while the latter uses predefined templates for producing captions. The retrieval-based models are further categorized into two categories: visual space and multi-modal space. The visual space models rank existing images based on visual similarity only whereas the multi-modal approaches perform similarity comparison over both the visual and text modalities. The deep learning-based IC models consist of encoder-decoder models and attention-based models. Similar to retrieval models, encoder-decoder models are also classified as visual space encoder-decoder models and multi-modal encoder-decoder models based on whether they manipulate visual or multi-modal features. Attention-based Image Captioning (AIC) is a state-of-the-art IC approach in which the model learns to focus solely on the salient elements while creating captions. The three basic varieties of AIC are region-based attention, semantic attention, and hybrid attention.

With the advent of multilingual datasets, the linguistic arena of IC is also expanding. Captions are being produced not only in English but also in Chinese Li et al. (2016, 2019), Lan et al. (2017), Hindi Rathi (2020), Dhir et al. (2019); Mishra et al. (2021), Japanese Miyazaki and Shimizu (2016), Yoshikawa et al. (2017), and Punjabi Kaur et al. (2021), among others. The introduction of non-English languages in IC datasets is further expanding the application domain of IC. As depicted in Fig. 4, both English and non-English-based implementations of IC frameworks may be used to perform Generic-purpose Image Captioning (GIC), Medical Image Captioning (MIC), and Remote Sensing Image Captioning (RSIC). The GIC frameworks provide instructive descriptions for the naturally captured images. Medically relevant captions are produced by MIC frameworks. The captions generated by MIC frameworks aid experts and doctors in accurately and reliably assessing critical medical conditions. RSIC is the brand new application area for IC frameworks that generate captions for images collected by Unmanned Aerial Vehicles (UAVs). Since GIC datasets are accessible to the public, they have been translated into other languages. As a result, GIC frameworks are currently being developed in both English and non-English languages. However, medical and remote sensing datasets are either inaccessible to the public or have not yet been translated. Accordingly, the present implementations of MIC and RSIC only produce English captions. Nonetheless, attempts are underway to convert medical and remote sensing image captioning databases into languages other than English.

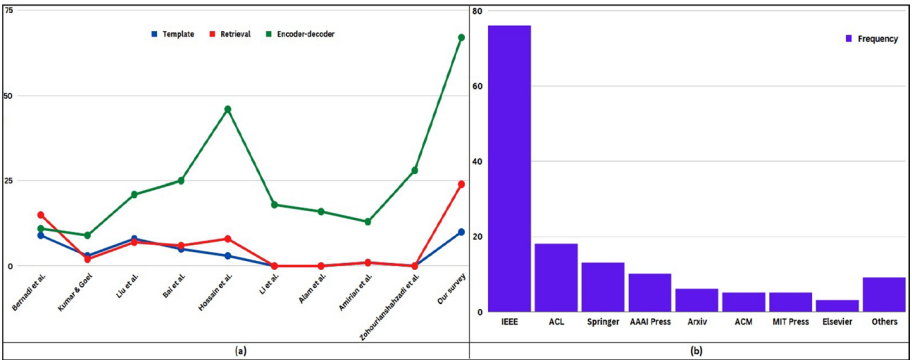

Fig. 2 Scope and coverage of articles: **a** based on architectures **b** based on publishers

Therefore, the non-English captioning frameworks will soon be utilized not just in these two application domains but also in a vast array of other application domains.

4 Machine learning-based approaches

Machine learning-based algorithms were used for both visual feature extraction and caption creation when IC initially emerged in 2010. Among the visual extraction strategies were the object detectors (Felzenszwalb et al. 2009), and global and local feature extractors like GIST, SIFT, HOG among others. All these visual extractors were utilized to extract the visual context of the image. Linguistic models such as statistical n-grams or a collection of templates were used for producing captions. Traditional models were component-oriented, with little information flow between sub-components. As a result, training such models was hard, and the captions generated were less reliable. The IC models that apply machine learning approaches are classified into two types namely direct generation models and retrieval models, based on whether they explicitly generate descriptions or reuse the existing captions. In Sects. 4.1 and 4.2, direct generation techniques and retrieval-based approaches are addressed.

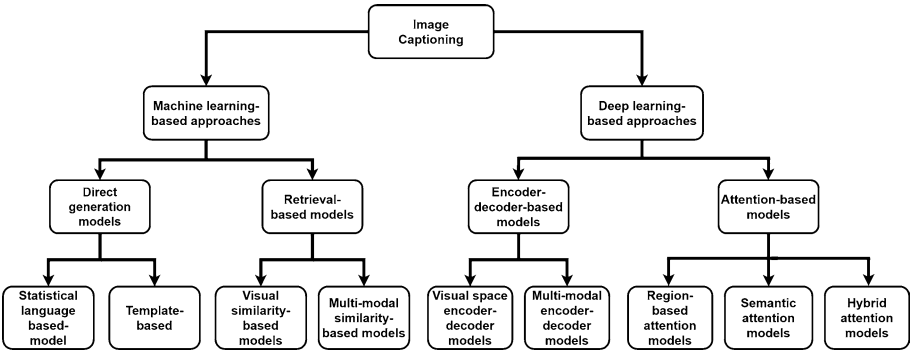

Fig. 3 A comprehensive taxonomy of image captioning based on their architecture

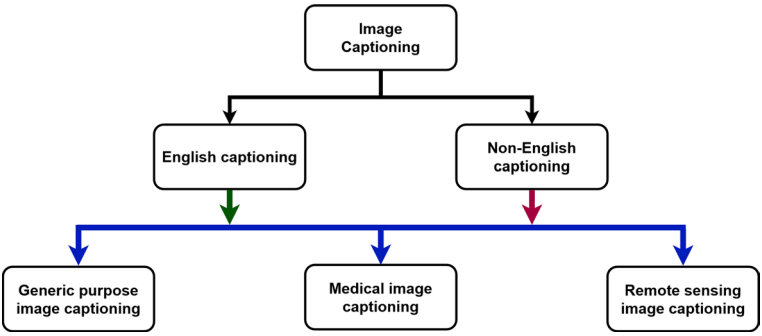

Fig. 4 An application-specific taxonomy of image captioning

4.1 Direct generation models

The earliest direct generation-based image captioning model named "BabyTalk" is proposed by Kulkarni et al. (2013). The model generates image descriptions by performing a series of steps namely content planning and surface realization. Content planning refers to the visual detection component and consists of Felzenszwalb et al. (2009) object detector, a linear SVM trained on low-level image features as a stuff detector, and a prepositional function to build spatial relationships around the detected objects. The surface realization is a language processing task that constructs a Conditional Random Field (CRF) using the former visual information. The language information inferred from CRF is processed by a language model such as n-grams or fitted into templates to generate the captions. Li et al. (2011) proposed a similar image captioning model utilizing a novel surface realization technique. It consists of a Radial Basis Function (RBF) kernel SVM to act as an attribute detector in addition to the existing detectors. The visual information is represented by the meaning tuple of *((adj1, obj1), preposition, (adj2, obj2))*. The model uses web-scale n-grams data trained over a training corpus of Wikipedia pages for content realization. The candidate phrases that best describe the image content are used to construct novel image descriptions.

Yang et al. (2011) implemented a language aid-based visual component built on the hypothesis that the output probability distributions of detectors are noisy and need additional linguistic inputs to make classification clear. It uses a GIST-based scene descriptor trained on SVM to detect scenes. The visual context of the image is represented using a quadruple of the form *(noun, verb, scene, preposition)* as shown in Fig. 5. A language model trained over the English Gigaword corpus is used to predict the possible verbs and the spatial relations between the objects. The model generates a sentence description by framing a dynamic programming problem, which is solved using a Hidden Markov Model (HMM). Mitchell et al. (2012) also introduced a similar language assistance-based model named "Midge" that leverages the syntactic word co-occurrence statistics to filter out the noisy output from the visual detectors. The model consists of three phases namely content determination, micro-planning, and surface realization. In micro-planning, the initial probability's outputs from the visual detectors are utilized to filter out the practical visual contents. Syntactic trees are built around the detected visual contents using their attributes, actions, and relations. The surface realization step at the end chooses a single tree

from the set of all possible generated trees to generate captions. Elliott and Keller (2013) proposed that the information about spatial relationships of image regions can help language models generate more human-like descriptions. They developed an image captioning model that uses a Visual Dependency Representation (VDR) to capture the spatial information between objects present in different image regions. The input query image is annotated using the LabelMe annotation tool (Russell et al. 2008). The visual information along with the VDR is processed by a template-based language generation model to produce the required image captions. Elliott and de Vries (2015) extended this work by introducing an approach to training the VDR parsing model without human supervision. The model detects the visual context of the image using a Regional Convolutional Neural Network (RCNN). The objects detected using RCNN are input to a VDR parsing model that predicts the spatial relations between different objects. A template-based language model leverages VDR and visual context information to generate a possible set of descriptions for the query image. The language model assigns a score to the generated description based on the training corpus evidence. The top-scored description is used as a caption for the query image.

Kuznetsova et al. (2014) use visually similar phrases retrieved from the dataset to implement a tree composition-based generation model. The model consists of a generalization step to remove any unproductive information. Noun phrases are extracted based on color, texture, and shape-based visual similarity. Verb phrases are extracted with the help of noun phrases and the prepositional and scene-based phrases are extracted using visual and spatial similarity. Generation of image description using retrieved phrases is framed as a constraint optimization problem solved using dynamic programming. Yatskar et al. (2014) retrofit the existing (Mitchell et al. 2012) model into a feature norm-based generation model. The dense annotation of image entities is called the feature norm. The latent variables that align phrases to objects and features are used to densely annotate the query image. The caption generation first distributes the existing annotated content followed by a re-ranking based on the generative model score and length of generated captions. The top-ranked sentences are used as the caption for the query image. Fang et al. (2015) proposed a deep learning-based direct generative model that uses Multiple Instances Learning (MIL) to train word detectors for eliminating any bias when detecting objects. The word distributions extracted from the visual features are input to a language model based on maximum entropy to generate image captions. The generated captions are further re-ranked using deep learning-based multimodal similarity.

Lin et al. (2015) proposed a dense captioning framework for indoor images. The model is built using a three-component architecture namely visual parser, generative grammar, and a text generation algorithm. The visual parser generates a scene graph for all the recognized visual content and attributes. The visual details contained in the scene graph are translated into a semantic tree. The generative grammar interprets these semantic trees as descriptive sentences. Table 3 compares direct generation-based IC techniques involving a range of characteristics.

4.2 Retrieval-based models

Retrieval-based IC, also known as ranking-based IC, is predicated on the assumption that images with equivalent visual contents have similar descriptions. The retrieval algorithms compare the incoming query image to the existing training images and rank them based on the degree of similarity. The visual or multi-modal features are used to compare

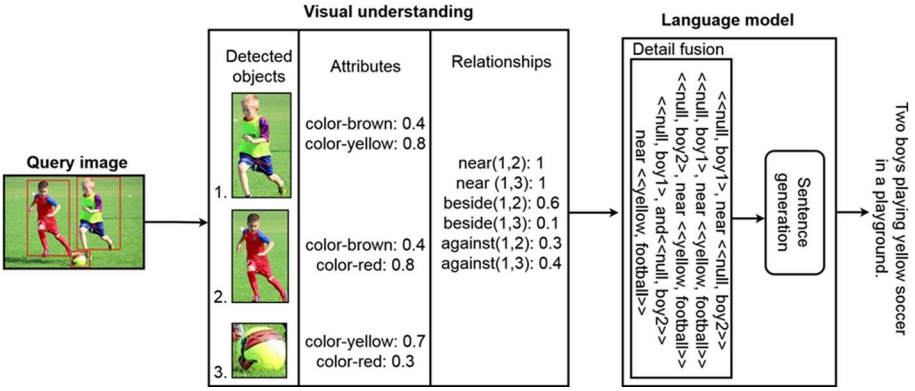

Fig. 5 The architecture of a direct generation model (Yang et al. 2011)

similarities. To implement the retrieval process, retrieval-based models employ an intermediate meaning space.

Farhadi et al. (2010) initialize an intermediary meaning space into which images and phrases are projected. Felzenszwalb et al. (2009) object detector and a GIST-based scene classifier extract feature from the query image. The image and sentence representations are projected into the multi-modal meaning space for similarity comparisons. The semantically related phrases are ranked higher and used as a query image caption. Ordonez et al. (2011) established a more robust model trained on a massive set of web-scale images. The model compares global and local similarity, and the sentences that rank higher are used as captions for the query image. Gupta et al. (2012) emphasized that the *(object, action, scene)* formulation cannot describe all image details adequately. To solve this constraint, they suggested a hybrid IC model that combines retrieval and generation approaches. The model creates a triplet of form *(((attribute; object); verb); (verb; prep; (attribute; object)); object; prep; object))*. Using a generation-based language model, this triplet can yield several descriptions. Kuznetsova et al. (2012) suggested a hybrid system based on Integer Linear Programming (ILP). The algorithm extracts four types of phrases (noun, verb, stuff, and scene phrases) from several candidate descriptions. The generation of descriptions from these words is framed as a constraint optimization problem, which is solved with ILP.

Hodosh et al. (2013) employ Kernel Canonical Correlation Analysis (KCAA) to link text and images in a shared induced space (Z) as shown in Fig. 6. The Canonical Correlation Analysis (CAA) compares a set of variables that are related. Images (I) and sentences (S) are the two sets in the case of IC. Both modalities are projected into *Z* where the semantically similar images and sentences are maximally correlated. Gong et al. (2014) used transfer learning to create a multimodal approach. CNN replaces the earlier feature extractors and a novel technique, Stack-Auxiliary Embedding (SAE), examines image and text features in a multimodal context. Karpathy et al. (2014) follow a similar multimodal deep learning approach. The model extracts sentence fragments and learns their multimodal associations. The model utilizes a global ranking objective function that analyses the similarity score of individual fragments of both modalities to compute the similarity between images and phrases.

Patterson et al. (2014) sort existing images based on global features derived with GIST, HOG, and spatial pyramids. Following that, the image features are split into

super-pixels and compared to existing visual classes. To compare words with images in multimodal space, a binary-Inverse Document Frequency (b-IDF) word representation of sentences is constructed. Mason and Charniak (2014) devised a non-parametric density-based estimating approach to minimize the noise in visual detector estimations. Using Euclidean distance, the model examines the similarity of multi-modalities. Verma and Jawahar (2014) designed a multimodal that generates the sentence's probability distribution vector and calculates the correlation between the multi-modalities using CCA and structured SVM.

Socher et al. (2014) developed a Dependency Tree-Recurrent Neural Network (DT-RNN) for capturing compositional semantic meanings. The model employs a CNN to extract visual attributes and maps each word to a d-dimensional vector. A deep learning-based max-margin objective function is used to learn the correlations between text and images. Chen and Zitnick (2015) developed a deep bidirectional multimodal system that performs both image retrieval and IC. The model uses CNN to extract visual information and RNN to perform multimodal computations for producing caption. Jeff et al. (2017) introduced Long-term Recurrent Convolutional Networks (LCRNs) that manage both spatial and temporal dependencies. For activity identification and caption creation, the model employs an end-to-end combination of CNN and LSTM. Devlin et al. (2015) used a pipelined as well as an end-to-end multimodal approach. In the pipelined model, the Maximum Entropy Language Model (MELM) leverages words recognized by CNN to construct descriptive phrases. The activation value of the penultimate layer is fed as input to the Multimodal Recurrent Neural Network (MRNN) to initialize its state. The model predicts captions word by word based on the hidden states.

Jia et al. (2015) developed gLSTM, a novel extension of LSTM, to generate captions that are closely connected to the query image. The gLSTM uses global semantic information to seed the hidden states for synthesizing the words that describe the input image. Karpathy and Fei-Fei (2014) developed a deep learning model based on dense captioning that produces descriptions for each component of the image separately. The RCNN is used to build a representation of the visual context. A Bidirectional Recurrent Neural Network (BRNN) is used to encode the training phrases into a fixed representation. To correlate visual representations with text, a multimodal space is trained using a max-margin objective function. Instead of a sophisticated recurrent language model, (Lebret et al. 2015) presented an IC model based on phrase association. The model trains a bilinear multimodal space to associate visual characteristics with a set of relevant sentences. To generate captions, the sentences are combined with a generative statistical language model. Lin et al. (2015) used an IC architecture that includes a visual parser, generative grammar, and a text-producing algorithm to characterize the complex indoor images. The visual parser extracts the characteristics of the input image and builds a scene graph using feature extractors such as RGB histograms, SIFT, and others. The scene graph is divided into many semantic trees, from which the text generation algorithm constructs captions.

To learn the correlation between the multi-modalities, (Mao et al. 2018) employed a log-likelihood function. An RNN is fed the visual context retrieved, that generates a description for the query image. To generate descriptions for clip-art images, (Gilberto et al. 2015) use abstract scenes and a Statistical Mechanical Translation (SMT) model. The model uses a Visual Dependency Grammar (VDG) to represent spatial relationships. To convey the major elements in the description, an ILP method is employed. Lebret et al. (2015) employ a probabilistic multimodal architecture to associate visual properties with a collection of phrases. Ushiku et al. (2015) introduce a new phrase learning approach called

Table 3 A table comparing direct generation-based image captioning methods

Year	Author(s)	Generation approach	Visual model	Language model
2011	Kulkarni et al. (2013)	Templates,Statistical model	Felzenszwalb et al. (2009) detector, Linear SVM	n-grams, Templates
2011	Li et al. (2011)	Templates, Statistical model	Felzenszwalb et al. (2009) detector, Linear SVM, RBF Kernel	Web-scale n-grams
2011	Yang et al. (2011)	Templates	Felzenszwalb et al. (2009) detector, GIST	Hidden Markov Model
2012	Mitchell et al. (2012)	Statistical model	Felzenszwalb et al. (2009) detector, SVM	Syntactic trees.
2013	Elliott and Keller (2013)	Templates	LabelMe annotator	Templates
2014	Kuznetsova et al. (2014)	Statistical model	Felzenszwalb et al. (2009) detector.	Phrase trees
2014	Yatskar et al. (2014)	Statistical model	LabelMe annotator, Feature norm	Probability distribution model
2015	Elliott and de Vries (2015)	Templates	RCNN, LabelMe annotator	Templates
2015	Fang et al. (2015)	Statistical model	AlexNet, VGG-Net,	Maximum entropy language model
2015	Lin et al. (2015)	Statistical model	3D RGB-D object detector	Syntactic trees.

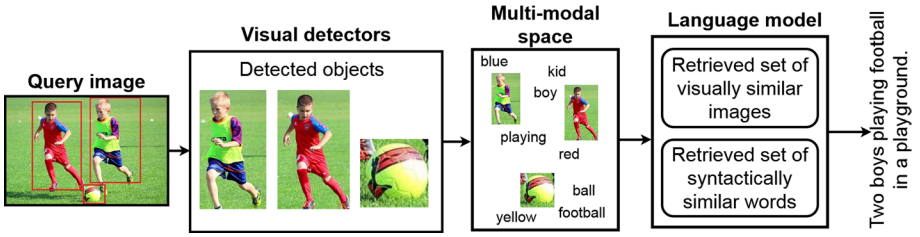

Fig. 6 The architecture of a retrieval-based model (Hodosh et al. 2013)

Common Subspace of Models and Similarity (CoSMoS). The strategy solves the problem of insufficient training samples while also reducing the complexity of multi-modality. The retrieval-based IC models are summarized in Table 4 based on different parameters.

5 Deep learning-based approaches

The machine-learning-based IC, discussed in Sect. 4, generates reasonable results on small datasets. However, because of the component-based design, performance degrades with complexity. Visual components in the template and retrieval-based models are trained to identify a few limited aspects. As a result, the system becomes biased, identifying just a subset of the enormous visual input. The use of hand-engineered feature extraction introduces another constraint. Since attributes are handcrafted, the emphasis is on feature engineering rather than genuine multi-modality training. Furthermore, template-based models are frequently difficult to analyze since the caption structures are predefined. Moreover, as the number of training samples and objects increases, the multimodal space gets more complicated. The deep learning-based IC overcomes these limitations by extracting features implicitly and handling complexities better than traditional approaches. Additionally, deep learning-based object identification models such as CNN and sequential models such as RNN and LSTM produce accurate results when detecting objects and producing captions. Based on their architecture, deep learning-based IC models are classified into two kinds namely encoder-decoder models and attention-based models. We did not investigate alternative deep learning models, such as compositional architectures, hybrid architectures, and others, to keep the scope of this research limited.

5.1 Encoder-decoder-based models

Machine translation systems influenced the encoder-decoder-based IC architecture. A machine translation system transforms a text sequence of one language into another. In these systems, a text sequence is encoded and decoded using an RNN. Inspired by this architecture, IC is also formulated as an encoder-decoder-based translation task. Both the encoder and decoder are trained jointly, and hence, these models are also called end-to-end models. Based on how these models learn multimodal correlations, they are divided into two categories namely multimodal encoder-decoders and visual space encoder-decoders. We explain both these kinds in Sects. 5.1.1 and 5.1.2.

5.1.1 Multimodal encoder-decoder techniques

Multimodal encoder-decoders are an extension of retrieval methodologies in which neural models such as CNN and RNN encode and decode images. The overall design of multimodal encoder-decoder consists of four components: an encoder that includes both CNN and RNN, a multimodal space, and a decoder that incorporates sequential models such as RNN or LSTM.

The earlier work is done by Kiros et al. (2014). They use a multimodal logarithmic bilinear function to predict word sequences conditioned on context vector as shown in Fig. 7. The logarithmic bilinear model can be considered a feed-forward neural network with a single hidden layer. Each word *(w)* of the vocabulary is represented using a D dimensional vector. The vocabulary set can be represented as $\{r_{w1}, r_{w2} \ldots, r_{wk}\}$ where $r_{wi} \in R^D$. If an input sequence of words $(w_1, w_2, \ldots, w_{n-1})$ is given, the n^{th} word prediction is calculated as illustrated in Eq. 1.

$$r_{wn} = \sum_{i=1}^{n-1} C^{(i)} r_{wi} \tag{1}$$

where C^i is $D \times D$ dimensional context matrix calculated over the previous *n-1* words of the caption. To condition the predicted word over the visual context,they proposed two approaches namely Modality-based Log Bilinear model (MLBL-B) and Factored 3-way Log Bilinear Model (MLBL-F). Kiros et al. (2014) expanded on their work by developing the Structured Context Vector Neural Language Model (SC-NLM), which decodes multimodal representations of images and texts (Kiros et al. 2014). The image encoding is also improved by using VGG Net as an image encoder.

5.1.2 Visual space encoder-decoder techniques

The multimodal encoder-decoder models perform better than retrieval systems regarding ranking and IC. However, they cannot model long-term visual-language interactions. To overcome such constraints, visual space encoder-decoder approaches have been developed. In visual space encoder-decoder systems, the features are not projected into a multimodal space. Instead, the vector representations are passed to the language model separately. The general architecture of visual space encoder-decoder consists of two components namely image encoder and decoder.

Vinyals et al. (2015) proposed an encoder-decoder model known as "Neural Image Caption". The model directly maximizes the probability of associating the correct sentence description with the query image by updating the framework parameters (θ) using the relation shown in Eq. 2.

$$\theta^* = argmax_\theta \sum_{(I,S)} logp(S||I;\theta) \tag{2}$$

where I and S represent the *(image, caption)* training pair and θ^* represents the modified parameters of the framework after maximizing over the current training batch.

Wang et al. (2018) proposed a bidirectional LSTM based IC model. Wang et al. (2017) later expanded their work by presenting an IC system that creates captions in two phases. The skeleton sentence is constructed in the first step by recognizing objects in the image. The identified items are revisited in the second step, and their attributes are

Table 4 A table comparing retrieval-based image captioning methods

Year	Author(s)	Retrieval approach	Visual model	Parser/ Language model
2010	Farhadi et al. (2010)	Multimodal	Felzenszwalb et al. (2009)detector, GIST	Curran & Clark parser
2011	Ordonez et al. (2011)	Multimodal	GIST, HOG, RBF kernal	-
2012	Gupta et al. (2012)	Hybrid	GIST, SIFT, Haar descriptor	Stanford CoreNLP toolkit
2012	Kuznetsova et al. (2012)	Hybrid	Felzenszwalb et al. (2009) detector, Scene and Stuff classifier	Berkeley PCFG parser
2013	Hodosh et al. (2013)	Multimodal	SIFT	BOW , Tagrank, and Trigram
2014	Gong et al. (2014)	Multimodal	CNN	WordNet
2014	Karpathy et al. (2014)	Multimodal	RCNN	Stanford CoreNLP toolkit
2014	Patterson et al. (2014)	Multimodal	GIST, Spatial pyramid, Color histogram	Binary-idf
2014	Mason and Charniak (2014)	Multimodal	GIST, Spatial pyramid, Color histogram	Real valued representation
2014	Socher et al. (2014)	Multimodal	CNN	Dependency tree RNN
2014	Verma and Jawahar (2014)	Multimodal	SIFT	Probability distributions
2015	Chen and Zitnick (2015)	Multimodal	VGGNet	Stanford CoreNLP toolkit, RNN
2015	Jeff et al. (2017)	Multimodal	AlexNet	LCRN
2015	Devlin et al. (2015)	Multimodal	VGGNet	LSTM, MRNN
2015	Jia et al. (2015)	Multimodal	VGGNet	NLTK toolbox, tf-idf, BOW
2015	Karpathy and Fei-Fei (2014)	Multimodal	RCNN	Bidirectiional RNN, MRNN
2015	Lebret et al. (2015)	Hybrid	VGGNet	SENNA, trigram
2015	Lin et al. (2015)	Visual model	RGB Histograms, SIFT	Stanford parser, Generative grammar
2015	Mao et al. (2018)	Multimodal	AlexNet,VGGNet	mRNN
2015	Gilberto et al. (2015)	Hybrid	VDG	Stanford parser, SMT
2015	Lebret et al. (2015)	Multimodal	VGGNet	SENNA, trigram
2015	Ushiku et al. (2015)	Multimodal	SIFT, AlexNet, VGGNet	Stop word filter

produced. Ren et al. (2017) proposed a reinforcement learning-based encoder-decoder model. The model has two subcomponents: policy network and value network. The former serves as the local guide, and the latter serves as the global guide. The sub-components are aligned to predict the description sentences close to the ground truth captions. Dai et al. (2017) proposed an image captioning system based on Conditional Generative Adversarial Networks (CGAN). The model is divided into two sub-models: generative subcomponent and evaluation subcomponent. The former creates query images description phrases. The latter evaluates the generated sentences based on the semantic similarity with the query image. Shetty et al. (2017) proposed a similar GAN-based IC model that generates multiple captions for a given query image. Liu et al. (2019) also proposed a Reinforcement Learning (RL) based IC model that uses policy gradient techniques instead of the traditional log-likelihood model to associate images with the captions.

The traditional encoder-decoder model uses LSTM based language models. LSTM cannot save visual information when generating long sentence descriptions. To overcome this limitation, (Gu et al. 2017) proposed to replace the LSTM-based language model with CNN. The CNN-CNN-based encoder-decoder model produces better results than the CNN-LSTM based models as they do not lose visual context while generating longer sentence descriptions. Aneja et al. (2018) and Wang and Chan (2018) further improved the CNN-CNN IC framework by introducing an attention mechanism.

Jie et al. (2021) discovered that the objective function utilized by present IC models conceals critical information to balance the error rates in training samples. They, therefore, introduce a unique global and local discrimination objective function for producing more precise captions. The innovative objective function maintains a balance between repressed and salient information. Lingxiang et al. (2021) upgraded the visual framework of the IC model by employing a Graph Convolutional Network (GCN) to recognize visual context. The GCN incorporates both regional and grid-level information that may be used to learn regional and global contexts. They also added a noise module, encouraging RNN to generate more descriptive and relevant captions. Yang et al. (2021) also adopt a structural development in which the LSTM inputs are adjusted to make the hidden states rely only on the visual context. As a result, the model eliminates any bias if the last word created is predicted unrelated to the image. Zhang et al. (2021) added a linguistic update to the IC system by adding parts of speech information into the IC model. The model applies some grammatical rules to determine the next element of speech using which next word of the caption is predicted. The algorithm creates the next word of the caption based on the visual context and previously produced caption and parts of speech information. Zhao et al. (2021) presented a method to use an IC model in cross-domain. The knowledge of the model trained over existing modalities

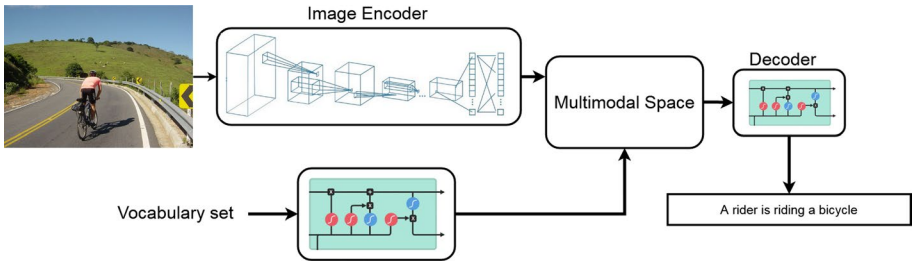

Fig. 7 The architecture of an encoder-decoder-based model (Kiros et al. 2014)

and domains is transferred to the current domain using generalization and transfer learning techniques.

Several languages have no literary transcriptions and are spoken only. On the other hand, the IC model interprets knowledge associated with image regions with text. As a result, the present IC models are constrained by the requirement for textual transcriptions of the target caption language. Effendi et al. (2021) proposed building a description system that learns to describe the visuals in terms of audio output to tackle this complex challenge. Hossain et al. (2021) use synthetic data generation approaches to achieve more generic training of IC models. They produce synthetic images with a Generative Adversarial Network (GAN) and synthetic captions with an attention-based approach. The model trains itself to match relevant phrases with visuals of interest. Jiang et al. (2021) improved both the encoder and decoder components of the IC model by using a pre-layer norm transformer module on the encoder side to improve image features and a multi-gate attention block. Kastner et al. (2021) developed an innovative IC approach that produces length-controlled captions. The model also creates captions with quality control that the user can set. Using data augmentation techniques, the model limits the caption lengths by substituting customized terms with generic phrases. Li et al. (2021) focused on the traffic-description domain and built an IC model that correctly detected all kinds of traffic objects and sceneries to make driving more interactive. The model analyses traffic scenes and generates suggestions for driving. The quantity of information collected from an image using IC models is modest, but it can be expanded. Min et al. (2021) investigated this IC area and developed a short tale IC approach. The model learns visual semantic alignment before creating skip-thought vectors from various stories. After that, the skip thought vectors are transformed into storyline sentences. Almost all present IC models solely use images from digital cameras. On the other hand, (Jing and Li 2021) used a unique IC model to interpret optical Scanning Electron Microscope (SEM) images. The model learns to explain and produce explanations for the patterns visible in SEM images.

Yumeng et al. (2021) introduce a news captioning approach for creating individual-specific news captions. The model employs a Graph Neural Network (GNN) to comprehend the semantic context and create captions. The approach comprises news templates in which generic terms are later replaced by specific words such as individual and location names found in the images. Zhou et al. (2021) similarly improve on the produced caption. The model substitutes the LSTM with a Gated Recurrent Unit (GRU), significantly improving caption quality. Bin et al. (2022) created an entity slot filling-based IC model using an image and a description that included some vacant slots to be filled.

Liu et al. (2021) suggest a RL-based IC model in which a credit assignment-based score is awarded to each vocabulary word at every generation step. The created caption is seen as an agent, while the external data is the environment. Ben et al. (2022) proposed a self-learning architecture for IC that uses unpaired image and text descriptions to learn. The model operates in two parts, first generating the pseudo-image-sentence pairings and then optimizing the samples for the objective function. A comparative analysis of encoder-decoder models is illustrated in Table 5.

5.2 Attention-based models

The encoder-decoder IC techniques adhere to the traditional machine translation design. While the classical technique structure works wonderfully for short translation sentences, the accuracy decreases as the sentence length grows. A similar problem is seen

in IC systems. If the quantity of information in the image is modest, the model creates flawless captions. However, if the image is complicated, the produced captions lose sight of visual context. The incorporation of the attention mechanism in IC alleviates the above-mentioned problem. AIC systems concentrate exclusively on salient information, ignoring the non-salient surroundings. Consequently, even if the image is complex or comprehensive, the model generates captions that just describe the salient contents. As a result, the AIC models produce more reasonable captions than earlier models. The attention mechanism may be introduced in a variety of ways, but the three most common are region-based attention, semantic attention, and hybrid attention.

5.2.1 Region-based attention techniques

The primary attention mechanism introduced in IC is region-based attention. Each query image is partitioned into equally sized image areas and at each time frame, the attention mechanism selects a new region to be attended. Xu et al. (2015) introduced region-based attention in IC. The model differs from the classical encoder-decoder models in two major ways. First, instead of extracting a single feature vector, a total of L vectors are extracted from query image. Second, the model uses an attention-based function (ϕ) to compute the context vector ($\hat{z}$) based on which RNN predicts the next word. Fig. 8 depicts the architecture of region-based attention framework Xu et al. (2015).

Consider $\{s_1, s_2, \ldots, s_n\}$ be the image regions to be attended by the attention function (θ) while generating the t^{th} word. The context vector of the attended region is computed using the following two types of proposed attention mechanisms.

i. *Hard attention:* In hard attention, only a single image region is attended at any given time frame. The context vector ($\hat{z}$) is dynamically updated as illustrated in Eq. 3.

$$\hat{z}_t = \sum_i s_{t,i}(a_i) \tag{3}$$

where $s_{t,i}$ is the region to be attended while generating the t^{th} word of the caption. The $s_{t,i}$ vector is one hot vector for hard attention as only a single image region is attended at each time frame.

ii. *Soft attention:* In soft attention, more than a single image region is attended at a given time frame. The context vector ($\hat{z}$) is dynamically updated as illustrated in Eq. 4.

$$\hat{z}_t = \sum_{i=1}^{L} \alpha_{t,i}(a_i) \tag{4}$$

where $\alpha_{t,i}$ is determined by considering all previously attended image regions $(s_{1,i}, s_{2,i}, \ldots, s_{t-1,i})$ in addition to the region presently selected.

Pedersoli et al. (2017) created a similar AIC approach that establishes an association between an image and semantic phrases. Liu et al. (2017a) developed a quantitative evaluation score-based AIC technique that focuses attention at many regional levels.

5.2.2 Semantic attention techniques

Semantic attention is a more advanced variant of region-based attention. Instead of splitting an image into regions, semantic attention attends to a collection of non-regular semantic

Table 5 A table comparing encoder-decoder-based image captioning methods

Year	Author(s)	Architecture	Visual model	Language model	Optimizer
2014	Kiros et al. (2014)	Multimodal	Alex Net	LBL	–
2014	Kiros et al. (2014)	Multimodal	Alex Net, VGG Net	LSTM, SC-NLM	–
2015	Vinyals et al. (2015)	Visual space	Google Net	LSTM	–
2016	Wang et al. (2018)	Visual space	Alex Net, VGG Net	LSTM	–
2017	Wang et al. (2017)	Visual space	Res Net	LSTM	–
2017	Ren et al. (2017)	Visual space	VGG Net	LSTM	–
2017	Dai et al. (2017)	Visual space	VGG Net	LSTM	–
2017	Shetty et al. (2017)	Visual space	Google Net	LSTM	–
2017	Liu et al. (2017b)	Visual space	Inecption-V3	LSTM	–
2017	Gu et al. (2017)	Visual space	VGG Net	Language CNN, LSTM	Adam
2018	Aneja et al. (2018)	Visual space	VGG Net	Language CNN	RMSProp
2018	Wang and Chan (2018)	Visual space	VGG Net	Language CNN	Adam
2020	Jie et al. (2021)	Visual Space	ResNet	LSTM	Adam
2020	Lingxiang et al. (2021)	Visual space	ResNet, Faster R-CNN	LSTM	Adam
2020	Yang et al. (2021)	Visual space	ResNet	LSTM	Adam
2020	Zhang et al. (2021)	Visual space	ResNet	LSTM	Adam
2020	Zhao et al. (2021)	Visual space	ResNet	LSTM	Adam
2021	Ben et al. (2022)	Visual space	RCNN	LSTM	Adam
2021	Bin et al. (2022)	Visual space	ResNet, Faster R-CNN, LSTM	LSTM	Adam
2021	Liu et al. (2021)	Visual space	ResNet	LSTM	Adam
2021	Zhou et al. (2021)	Visual space	VGG-16	Bi-GRU	Adam
2021	Yumeng et al. (2021)	Visual space	VGG-19	LSTM, Templates	Adam
2021	Jing and Li (2021)	Visual space	VGG-16	LSTM	SGD
2021	Min et al. (2021)	Visual space	RCNN	GRU	Adam
2021	Li et al. (2021)	Visual space	VGG-16	LSTM	SGD
2021	Kastner et al. (2021)	Visual space	Faster R-CNN	BERT	–
2021	Jiang et al. (2021)	Visual space	Faster R-CNN	Transformer	Adam
2021	Hossain et al. (2018)	Visual space	DenseNet	Bi-LSTM	Adam
2021	Haque et al. (2021)	Visual space	Parallelized capsule network	Transformer	Adam
2021	Effendi et al. (2021)	Visual space	ResNet	Transformer	Adam

areas. Attending such semantic regions makes the AIC system more precise and reliable while discussing salient contents. Jin et al. (2015) introduced the semantic attention technique in IC model. The model captures scene specific context and attends it while generating captions. A CNN extract feature vector for each visual area. The global context vector is extracted in addition to the feature vector. To create captions, the feature vectors, and context vectors are supplied in the LSTM decoder. Lu et al. (2017) suggested that attention may not be required for generating all the words of the caption. As a result of this research, a model based on selective attention is offered. The model implicitly determines whether or not to pay attention to the word currently being generated. Gan et al. (2016) developed

a semantic attention-based compositional network in which the LSTM parameters are constructed using the attended semantic notions. Tavakoli et al. (2017) studied human scene description abilities and created a salience-based semantic attention model. A framework for visual question answering and image captioning is created by Qi et al. (2017) which directly learns high-level semantic structures.

Yao et al. (2018) developed an attention model based on GCN that incorporates semantic and spatial objects. The model first generates a graph of the observed visual contents and then develops a context vector using the GCN. The context vector is then fed to the LSTM decoder to generate captions. Lu et al. (2018) and Yang et al. (2019b) constructed a traditional slot filling-based attention technique, where semantic contents filled template slots for generating captions. Ke et al. (2019) recommended paying attention to the words to improve captioning. Huang et al. (2019) introduced a module that extends the traditional attention technique to identify the relevance between attention outcomes and queries. To improve captioning reliability, (Gao et al. 2019) presented a two-pass AIC model that leverages an intentional residual attention network.

Wang et al. (2019) and Yao et al. (2019) built a hierarchical attention-based IC model that computes attention at multiple levels. Herdade et al. (2019) developed an object-relational transformer that directly contains information about the spatial relations between items identified through attention. Pan et al. (2020) use a unified attention block that makes use of both spatial and channel-wise attention. Liu et al. (2019) suggested a global and local information exploration and distilling strategy for word selection that abstracts scenes, spatial data, and attribute level data. Zhou et al. (2020) provide improved multi-modeling using parts of speech.

5.2.3 Hybrid attention techniques

A hybrid attention model combines two or more attention types in a single framework. The hybrid attention makes a model more robust by leveraging the power of several attention techniques in a single framework. Learning a high-dimensional hybrid transformation matrix of the query image is proposed by Ye et al. (2018). The model implements hybrid attention techniques such as spatial, regional, and channel-wise attention using the transformation matrix. Qin et al. (2019) introduced the look back approach, which incorporates prior attention input into the current time frame. Yang et al. (2019a) also implemented a similar auto-encoder model that incorporates the language inductive bias into the encoder-decoder framework. Li et al. (2019) proposed an augmented transformer model to achieve

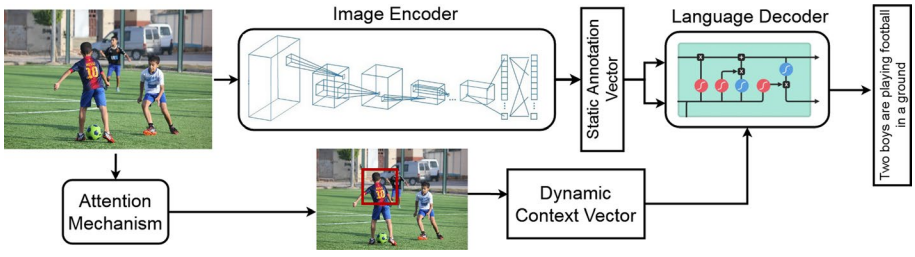

Fig. 8 The architecture of an attention-based model (Xu et al. 2015)

both vision-guided attention and concept-guided attention in a single framework. Wang et al. (2020) employed a memory mechanism-based attention method to simulate human visual interpretation and captioning skills. Sammani and Melas-Kyriazi (2020) proposed a unique IC strategy that uses the current training captions instead of creating novel captions from scratch. Jun et al. (2020) developed a multi-modal transformer for image captioning that captures both intra-modal and inter-modal interactions in a single attention block. Cornia et al. (2020) created a meshed memory transformer-based model for attention that employs both low and high-level visual information. Guo et al. (2020b) synthesize captions using a self-attention extension to overcome the limits of the transformer-based IC model. Wang and Gu (2021) went on to design a dynamic AIC model that generates captions without ignoring function words. Regional feature vectors extracted by object detection networks (CNNs) offer extensive information about local regions but lack the global context required to establish the association between distinct semantic regions. In contrast, the conventional grid-level features effectively maintain the global context. Luo et al. (2021) introduced a novel Dual-Level Collaborative Transformer (DLCT) that incorporates both global and local feature vectors while generating captions. Using a unique Dual-Way Self Attention module, the DLCT integrates the global and regional contexts to incorporate spatial information into the visual input. Zhang et al. (2021) developed a Grid-Augmented module that boosts the assimilation of spatial features into the visual input to effectively counteract the loss caused by the flattening of visual context. In addition, they incorporated an adaptive attention module to distinguish between visual and non-visual keywords for efficient caption production. Fang et al. (2022) proposed a visual detector-free IC model utilizing a vision transformer that directly acts on the grid features, streamlining and accelerating the training process. Moreover, a novel Concept Token Network is coupled with the language model to inject the semantic details into the caption. Wang et al. (2022) replaced the object detection encoder network with a SwinTransformer to extract the grid-level features from the input image. A refining encoder module refines the features before feeding them to the decoder for caption synthesis. Table 6 summarizes the AIC techniques over several attributes.

5.2.4 Non-autoregressive attention techniques

All the attention-based IC models discussed so far use an autoregressive captioning strategy meaning that the new token is conditioned on the previously generated word and the dynamically chosen image feature vector. Even though this paradigm gets good outcomes, it wastes a lot of computational resources by limiting concurrency. Motivated by the current paradigm of non-autoregressive machine translation frameworks that synthesize the whole translation simultaneously, the attention-based IC is also being framed using a similar methodology. Non-autoregressive attention-based models use a novel training paradigm in which the loss function is optimized to simultaneously learn the whole caption. Guo et al. (2020a) introduced one of the earliest attention-based non-autoregressive IC frameworks with a novel training approach of Counterfactual-Critical Multi-Agent Learning to optimize directly on captions rather than individual words. To improve the performance of their non-autoregressive model, they additionally incorporated vast amounts of unlabeled data. Fei (2021) argued that utilizing solely the non-autoregressive method results in

tokens that are duplicated or incomplete. As a consequence, they suggest a technique that is partially non-autoregressive and treats the caption as a series of concatenated word groupings. Each group formed is considered a separate entity and is thus processed simultaneously. Utilizing word groups combines the sequential aspect of the caption with the ability to work concurrently.

6 Application-specific image captioning

The potential uses of IC are diverse, necessitating a review of their application areas as well. General-purpose Image Captioning (GIC), Medical Image Captioning (MIC), and Remote Sensing Image Captioning (RSIC) are the three primary application areas of IC, as shown in Fig. 4. The GIC is so named because the images utilized in these approaches are often generic images of our surroundings, people, and animals. Furthermore, the produced descriptions are of little vital importance, and hence these models explain contents with flexibility. However, the MIC and RSIC are employed in critical and security applications. The GIC frameworks are currently being implemented in both English and non-English languages. All the models discussed thus far in Sects. 4 and 5 are examples of GIC frameworks producing English captions. The non-English-based implementations of GIC are discussed separately in Sect. 6.3 since its datasets differ from those of the English GIC frameworks. Additionally, the pre-processing stages, vocabulary size, and word tokens also vary between English and non-English captioning. As a result, a direct comparison between these two captioning categories is not possible.

6.1 Medical image captioning

Hou et al. (2021) created a chest x-ray report-generation using deep learning. The model is tuned for diagnostic accuracy and linguistic fluency, both lacking in prior efforts. The model employs a reinforcement learning-based training approach to learn the inter-modular relationships. Park et al. (2021) conduct a similar investigation to determine which feature representations and decoder algorithms are optimal for captioning medical x-ray images. They concluded that the transformer-based decoder outperformed LSTM approaches for captioning x-ray images. Table 7 summarizes the MIC approaches. MIC frameworks are presently only producing English captions since most MIC datasets are proprietary and unavailable to the public. In the future, however, medical databases may be translated into other languages, allowing the captioning of medical images in languages other than English.

6.2 Remote sensing image captioning

Hoxha et al. (2020) emphasized Remote Sensing (RS) images' visual properties to create a RSIC framework that generates accurate descriptions for the given query images. Cheng et al. (2021) leveraged attention mechanism to create a similar cross-domain multi-modal retrieval framework for RS images. Wang et al. (2020) proposed a retrieval

Table 6 A table comparing attention-based image captioning methods

Year	Author(s)	Attention type	Image encoder	Language decoder	Optimization
2015	Xu et al. (2015)	Region	VGG Net	LSTM	RMSProp, Adam
2015	Jin et al. (2015)	Semantic	VGG Net	LSTM	Adam
2017	Lu et al. (2017)	Semantic	ResNet	LSTM	Adam
2017	Gan et al. (2016)	Semantic	ResNet	LSTM	Adam
2017	Pedersoli et al. (2017)	Region	VGG Net	RNN	Adam
2017	Tavakoli et al. (2017)	Semantic	VGG Net	LSTM	–
2017	Liu et al. (2017a)	Region	VGG Net	LSTM	Adam
2017	Qi et al. (2017)	Semantic	VGG Net	LSTM	–
2018	Yao et al. (2018)	Semantic	RCNN	LSTM	Adam
2018	Ye et al. (2018)	Hybrid	ResNet	LSTM	Adam
2018	Lu et al. (2018)	Semantic	RCNN	LSTM	Adam
2019	Ke et al. (2019)	Semantic	RCNN	LSTM	–
2019	Qin et al. (2019)	Hybrid	RCNN	LSTM	Adam
2019	. Huang et al. (2019)	Semantic	RCNN	LSTM	Adam
2019	Wang et al. (2019)	Semantic	RCNN, ResNet	LSTM	Adam
2019	Yao et al. (2019)	Semantic	RCNN	LSTM	Adam
2019	Yang et al. (2019b)	Hybrid	ResNet	LSTM	Adam
2019	Wang et al. (2019)	Semantic	RCNN	LSTM	Adam
2019	Yang et al. (2019a)	Semantic	RCNN	LSTM	Adam
2019	Herdade et al. (2019)	Semantic	RCNN	Transformer	Adam
2019	Li et al. (2019)	Hybrid	VGG Net	Transformer	Adam
2020	Pan et al. (2020)	Semantic	RCNN, ResNet	LSTM	Adam
2020	Liu et al. (2019)	Semantic	RCNN	LSTM	Adam
2020	Wang et al. (2020)	Hybrid	RCNN	Bi-LSTM	Adam
2020	Sammani and Melas-Kyriazi (2020)	Hybrid	RCNN	LSTM	Adam
2020	Zhou et al. (2020)	Semantic	RCNN	LSTM	Adam
2020	Jun et al. (2020)	Hybrid	RCNN	Multimodal Transformer	Adam
2020	Cornia et al. (2020)	Hybrid	RCNN, ResNet	Transformer	Adam
2020	Guo et al. (2020b)	Hybrid	RCNN	Transformer	Adam
2020	Guo et al. (2020a)	Non-Autoregressive	RCNN	Transformer	Adam
2021	Wang and Gu (2021)	Hybrid	RCNN	LSTM	–
2020	Fei (2021)	Non-Autoregressive	RCNN	Transformer	–
2021	Luo et al. (2021)	Hybrid	RCNN	Transformer	Adam
2021	Zhang et al. (2021)	Hybrid	ResNet	Transformer	Adam

Table 6 (continued)

Year	Author(s)	Attention type	Image encoder	Language decoder	Optimization
2022	Fang et al. (2022)	Hybrid	Vision Transformer	Transformer	Adam
2022	Wang et al. (2022)	Hybrid	Swin Transformer	Transformer	Adam

topic-based RSIC technique in which the complicated job of RSIC is performed in two steps. First, the RS image topics are identified. Second, a descriptive phrase for the image is generated.

The variety and alignment scales of objects in different viewpoints are among the most severe challenges in RSIC. Huang et al. (2021) devised a de-noising approach for visual feature extraction to address this issue. According to Li et al. (2021), the cross-entropy objective function employed by earlier RSIC frameworks attempts to train the model to predict the following word with exact probability only. However, there may be a variety of synonyms that can be used instead of other terms-as a result, training the model to predict only a specific word causes the system to be overtrained. They offer a unique truncation cross-entropy objective function that avoids the overfitting problem to address this issue. Ma et al. (2021) used multiscale and multifeat attention in RSIC to realize the notion of attention-based IC in RS. Yuan et al. (2020) build a similar attention-based framework by introducing multi-head and multi-level attention to the RSIC framework using GCN. Sumbul et al. (2021) generated a unique summary of all the captions linked with the image, resulting in a single description with no repetitive information. As a result, the model only trains using the essential information offered in the captions. Wang et al. (2021) initiated the explain ability in RSIC's encoder-decoder structure for more meaningful descriptions. Table 8 summarizes the RSIC models over various parameters. We were unable to identify any RSIC framework capable of producing captions in non-English languages at the time this study was conducted. However, such RSIC frameworks may be developed in the near future.

6.3 Non-English image captioning

Researchers are now more focused on non-English image captioning. Captions in languages other than English are generated by the non-English IC models. The non-English IC discipline is now one of the most active research areas. Non-English IC systems are in high demand these days since many individuals do not understand English. As a result, non-English IC systems may be extremely beneficial when integrated with other application fields such as traffic description systems, news captioning, medical image captioning, and remote sensing image captioning. In the future, a significant amount

Table 7 A table comparing medical image captioning methods

Year	Author(s)	Type	Visual model	Language model	Optimizer
2021	Hou et al. (2021)	Medical report generation	ResNet	LSTM	Adam
2021	Park et al. (2021)	Medical image captioning	ResNet	Transformer	Adam

of research will indeed be conducted focusing on building frameworks and datasets for non-English IC.

Miyazaki and Shimizu (2016) constructed "YJ Caption 26K Dataset", the first non-English caption dataset. YJ Caption 26K is a Japanese version of the MS-COCO dataset in which captions are collected from crowdsourcing. To find the best way to describe an image in the Japanese language, (Miyazaki and Shimizu 2016) compared three learning techniques, i.e., monolingual, alternative, and transfer learning. Their research found that the transfer learning technique is the best way to generate a Japanese caption. Yoshikawa et al. (2017) developed "STAIR Caption", a Japanese caption dataset also based on MS-COCO. STAIR Caption is the largest Japanese image caption dataset. Yoshikawa et al. (2017) compared their model trained with the STAIR Caption dataset against machine-translated dataset. Tsutsui and Crandall (2017) used the YJ caption 26K dataset for multilingual image captioning.

To generate Chinese caption, (Li et al. 2016) collected image descriptions of Flicker 8K images from three sources, i.e., crowdsource, machine translator, and human translator. Li et al. (2016) found out that the model trained with machine-translated captions outperforms the model trained with the human-translated caption, and the model trained with crowdsource caption is at the top in terms of accuracy. Lan et al. (2017) proposed a "Fluency guided framework" where they developed fluent machine-translated image descriptions by editing non-fluent machine-translated sentences manually. Li et al. (2019) developed the "COCO-CN" dataset, the Chinese version of the MS-COCO dataset collected from crowdsourcing.

Other than Japanese and Chinese captioning, little research has been done to generate image descriptions in German (Elliott et al. Elliott et al. 2015)), Dutch (Miltenburg et al. van Miltenburg et al. 2017)), French, and Spanish language.

Kaur et al. (2021) implemented a deep learning-based Punjabi language image captioning system. The model consists of an encoder-decoder structure. VGG Net is used to extract image features, and the extracted features are fed to the LSTM states. The language model generates captions in the Punjabi language.

Rathi (2020) implemented a deep learning-based image captioning model for the Hindi language. Rathi (2020) first converted the English Flickr 8K dataset to Hindi using google translator. The model consists of an attention-based encoder-decoder model where a ResNet is used to extract image features. The extracted feature vectors are fed to the language model to generate captions in the Hindi language. Dhir et al. (2019) manually

Table 8 A table comparing remote sensing image captioning methods

Year	Author(s)	Visual model	Language model	Optimizer
2020	Hoxha et al. (2020)	ResNet	mRNN	–
2020	Huang et al. (2021)	VGG-16, ResNet	LSTM	Adam
2020	Li et al. (2021)	VGG-16, AlexNet, ResNet, GoogleNet	LSTM	Adam
2020	Ma et al. (2021)	ResNet	LSTM	Adam
2020	Sumbul et al. (2021)	ResNet, DenseNet	LSTM	SGD
2020	Wang et al. (2020)	VGG-16	LSTM	Adam
2020	Wang et al. (2021)	VGG-16, AlexNet, ResNet, GoogleNet	Transformer	Adam
2020	Yuan et al. (2020)	ResNet,GCN	LSTM	Adam
2021	Cheng et al. (2021)	ResNet,Inception V3	Bi-GRU	–

annotated the MS-COCO dataset to Hindi captions. Using this novel Hindi-based MS-COCO dataset, (Dhir et al. 2019) implemented an attention-based Hindi image captioning model. ResNet is used to extract image features, and LSTM based language model generates the Hindi image captions. Dhir et al. (2019) extended their initial work in Mishra et al. (2021) by implementing a transformer-based Hindi image captioning model.

7 Benchmark datasets

Image captioning, being one of the active research areas of artificial intelligence, is presently getting extensively investigated. Visual models, linguistic models, and multimodal couplings are all making inroads into the developmental study. The datasets on which IC models are trained are vital to their efficiency. Because the efficiency of deep learning approaches is proportional to the number of training instances, datasets are essential for robust model training. Several benchmark image captioning datasets are now available for training and evaluation. Both the quantity and quality of these databases have grown over time. Regularly, new and specialized datasets are delivered. The benchmark datasets used in image captioning are summarized in Table 9.

i. *IAPR TC-12:* IAPR TC-12 Grubinger et al. (2006) is a primarily gathered collection of around 20,000 images, with 1-5 captions for each image. The images depict a diverse visual array of people, animals, and landscapes, among others.
ii. *PASCAL 1K:* One of the early benchmark datasets used in image captioning was PASCAL 1K Rashtchian et al. (2010). The collection contains 1,000 images, each with five captions. The dataset is made up of lots of mixed visual scenes from PASCAL VOC that include humans, animals, and their surroundings.
iii. *SBU captioned dataset:* The images received in response to the input text queries on Flick.com produces the SBU captioned dataset Ordonez et al. (2011). A total of 1,000,000 images were obtained, which included a variety of visual scenarios. These retrieved images are filtered out, and only the precise images are used in the dataset. Each image is accompanied by one caption.
iv. *Flickr 8K:* Flick 8K Hodosh et al. (2013) is a collection of 8000 images of people and animals gathered from the website *flick.com*. There are five captions for each of these images. The training set contains 6,000 images, while the test and validation sets each contain 1,000 images.
v. *Flickr 30K:* Flickr 30K Gong et al. (2014) consists of 30,000 images downloaded from Flickr.com with 5 captions per image. There are a total of 158K captions, and 276,000 manually annotated bounding boxes matching each object.
v. *MS-COCO:* MS-COCO (Microsoft-Common Objects in Context) Lin et al. (2014) is the most often used benchmark image captioning dataset. The dataset contains 328,000 images of every day's complicated actions with a total of 91 item types. Each image has five captions. Several image captioning challenges are being run on this dataset.
vi. *Visual Genome:* The Visual Genome dataset Krishna et al. (2017) differs from the others in that it includes captions for each section or scene of the images. The dataset

contains about 108,249 images with 35 objects, 26 characteristics, and 21 pairwise relationships between items.

vii. *Instagram dataset:* Park et al. (2017) created and employed the Instagram Dataset in their IC model that predicts post description and hashtags. The dataset consists of 10,000 Instagram photographs, the majority of which are of celebrities. There are 1.1K post descriptions and 6.3K hashtags in the dataset.

8 Evaluation metrics

Many evaluation metrics are used to assess the performance of image captioning systems, such as comparing produced captions to reference captions based on n-gram sequences, semantic meaning, morphological similarity, and other factors. The following metrics are frequently employed in the evaluation of image captioning.

i. *BLEU:* The BLEU Papineni et al. (2002) (Bilingual Evaluation Understudy) metric is a modified precision metric used to assess the quality of the machine-translated text. Individual n-grams from the produced captions are compared to a collection of reference captions, and scores are computed. The BLEU score is calculated as the number of n-grams found in both produced and reference captions.

ii. *METEOR:* METEOR Banerjee and Lavie (2005) (Metric for Evaluation of Translation with Explicit Ordering) is a more sophisticated accuracy and recall-based metric. METEOR considers the morphology of words while matching n-grams. The metric computes unigram matches at multiple levels, including exact match, porter stem match, and synonyms. After calculating all such matches, the accuracy, recall, and F-Mean are determined as described in Eqs. 5, 6, and 7

Table 9 A table comparing benchmark image captioning datasets

Year	Author(s)	Dataset	Images	Captions	Scenes	Image source
2006	Grubinger et al. (2006)	IAPR TC-12	20000	1-5	Mixed	Primary collected
2010	Rashtchian et al. (2010)	PASCAL 1K	1000	5	Mixed	PASCAL VOC
2011	Ordonez et al. (2011)	SBU dataset	1000000	1	Mixed	Flickr.com
2013	Hodosh et al. (2013)	Flickr 8K	8092	1-5	People and animals	Flickr.com
2014	Gong et al. (2014)	Flickr 30K	31783	5	People and animals	Flickr.com
2014	Lin et al. (2014)	MS-COCO	328000	5	Mixed	Web
2017	Krishna et al. (2017)	Visual Genome	108249	1-5	Mixed	Web
2017	Park et al. (2017)	Instagram dataset	10000	–	People	Instagram

$$Precision(P) = \frac{Unigrams\ common\ in\ reference\ and\ system\ translation}{Total\ number\ of\ unigram\ in\ system\ translation} \tag{5}$$

$$Recall(R) = \frac{Number\ of\ unigramin\ system\ translation}{Total\ number\ of\ unigram\ in\ reference\ translation} \tag{6}$$

$$FMean = \frac{10PR}{R + 9P} \tag{7}$$

The METEOR is computed using the above-mentioned parameters as illustrated in Eq. 8.

$$METEOR = FMean * (1 - Penalty) \tag{8}$$

iii. *ROUGE:* ROUGE Lin (2004) (Recall Oriented Understudy for Gisting Evaluation) is a machine translation evaluation that is based on recall. ROUGE has several variants, including ROUGE-N (calculated over n-grams), ROUGE-L (longest common sub-sequence), and ROUGE-W.

iv. *CIDEr:* CIDEr Vedantam et al. (2015) (Consensus-based Image Description Evaluation) compares the resemblance of a produced caption to human-written ground truth captions. CIDEr takes into account the concepts of grammar, saliency, accuracy, and precision. CIDEr is computed as shown in Eq. 9.

$$CIDEr(c_i, S_i) = \sum_{n=1}^{N} w_n CIDEr_n(c_i, S_i) \tag{9}$$

where w_n is the n-gram size, c_i is the candidate caption produced by the IC framework and S_i indicates the set of ground truth captions for the query image.

v. *SPICE:* Because traditional n-gram assessment measures are susceptible to overlaps, SPICE Anderson et al. (2016) (Semantic Proportional Image Captioning Evaluation) is a unique image captioning metric. SPICE analyses captions by taking the semantic information of the image into account. SPICE creates a scene graph of both the reference and produced captions to match their semantic contents.

vi. *WMD:* The WMD Kusner et al. (2015) (Word Mover's Distance) measures the dissimilarity between the generated and the ground truth captions by exploiting the multimodal vector embedding spaces, as opposed to the other evaluation metrics that directly rely on n-gram matches. WMD is the minimum distance that the embedded word representations of one sentence must travel to reach the embedded word representations of another sentence. Suppose an IC framework with an n-dimensional multimodal vector embedding space is trained on a word vocabulary of size *'d'*. Let $P = \{p_1, p_2, p_3, \ldots, p_x\}$ and $G = \{g_1, g_2, g_3, \ldots, g_y\}$ be the predicted and ground truth captions respectively such that $\{P, G\} \in \mathbb{R}^{d*n}$. The overall dissimilarity between the generated and ground truth captions is computed using the Euclidean distance between the multimodal embeddings of the words as indicated in Eq. 10.

$$Dissimilarity(P, G) = \sum_{i=1}^{x} min(\sum_{j=1}^{y} (\|p_i^d - g_j^d\|^2)) \tag{10}$$

9 Open research challenges

Deep learning-based IC has advanced rapidly in recent years. Our examination shows that the IC field is still improving. The researchers must concentrate on the undeveloped aspects of IC for more efficient and reliable results. Although the IC is currently facing a variety of issues, the following are the most critical.

i. *High dimensionality and complexity:* Image captioning requires both image and text-based labeled data for training. Since the input training data is very complicated with several characteristics and features, the total complexity of the model is very high. Deep learning-based models, on the other hand, require a large dataset to train a model with high efficiency. However, with the increase in the training instances, the complexity increases further. Thus, one of the major challenges in image captioning is to increase the dataset size while keeping the model training overhead stable.
ii. *Inefficiency when generating extended sentences:* The inclusion of RNN in language models causes sequential information to degrade over time. LSTMs solved the vanishing and exploding gradient problems, however, a similar issue arises when creating prolonged captions. While generating lengthier descriptions, the semantic information contained in the context vector decays. The retention of semantic information while generating extended descriptions is another major research challenge of IC.
iii. *Limited accuracy and efficiency in real-time situations:* Image captioning, being a semi-developed field, has a lot of room for improvement in terms of efficiency and reliability of generated captions. There are a limited number of visual scenes and entities in the datasets that train IC models. As a result, when captioning real-time images, such IC systems usually provide erroneous descriptions. Vital applications, such as medical captioning and remote sensing captioning however need precise descriptions. As a result, one of the key research concerns in IC is training more robust models for real-time circumstances.
iv. *Attention misinterpretation:* IC attention mechanisms are still distant from the extraordinary human vision system. In many circumstances, the attention mechanism sees background features as prominent and attempts to focus more on them. As a result, the produced descriptions are unreliable since they describe irrelevant information. Thus, another important research issue in IC is to limit the attention misinterpretation to focus only on the salient information.
v. *Scarcity of application-specific datasets:* As per the review, there is a paucity of image captioning datasets. There are few reliable datasets available for general-purpose image captioning. However, the quantity of data available for training in application-specific domains, such as medical captioning (Hou et al. 2021; Park et al. 2021) and remote sensing captioning (Hoxha et al. 2020; Cheng et al. 2021; Wang et al. 2020), is currently insufficient. As a result, the research community must concentrate on the collection of such application-specific datasets that will assist future academics in developing efficient models.
vi. *Limited use of modern deep learning techniques:* Image captioning gradually evolved from handcrafted feature extraction-based processes to contemporary deep learning-based systems. The CNN and LSTM-based deep learning models are used in the

present encoder-decoder architecture of image captioning. While this model produces excellent results, combining it with more sophisticated and innovative deep learning approaches like transformers and graph convolutional networks produces more efficient outcomes. As a result, adopting these advanced deep learning techniques in IC remains a challenge for the future.

v. *Lack of non-English IC framework in MIC and RSIC:* Currently, the non-English IC is implemented exclusively in the GIC domain. However, such frameworks may be quite beneficial for assisting the general audience that does not comprehend English. Thus, implementing non-English captioning with reliable efficiency in various application domains such as medical and remote sensing is an open research challenge in the area of image captioning.

vi. *Scarcity of non-English image captioning datasets:* As mentioned in Sect. 6.3, there are merely a few datasets published for non-English languages such as Japanese, Chinese, Hindi, and Punjabi. However, since contemporary deep learning approaches are data-driven, they require more data for efficient training. Accordingly, expanding the current datasets and producing novel non-English IC datasets is a significant research obstacle that needs to be addressed in the future.

vii. *Deficiency of interlanguage knowledge transfer techniques:*English-based IC frameworks are revolutionary and produce trustworthy captions. However, non-English IC frameworks have low reliability. Therefore, researchers must develop techniques for information transfer so that the English-trained models may be utilized to train other language-based IC frameworks. Such solutions will further enhance the foundations for non-English captioning.

10 Conclusions

This study provides a comprehensive review of IC methodology as well as cutting-edge IC techniques. The major IC approaches are broadly classified based on the underlying architecture and application areas. Our research demonstrates that deep learning techniques contribute significantly to IC. This is because deep learning-based frameworks offer higher learning accuracy. Medical image captioning and non-English image captioning models are two of the domains that have received the least attention. Few models in non-English languages have been constructed, but they require additional modification to perform better. The concept of the attention mechanism, which enhances semantic attentiveness while producing captions, has recently been introduced. In the future, the IC may be investigated in additional application areas like security, robotics, and social media, to name a few. For more accurate image captioning, innovative deep learning approaches like transformers and GCN may be used.

Appendix

See Tables 10, 11, 12

Table 10 Comprehensive information about the publishers of the research articles that are cited in this survey

Author(s)	Journal/Conference	Publisher
Alam et al. (2020)	Neurocomputing	Elsevier
Amirian et al. (2020)	IEEE Access	IEEE
Anderson et al. (2016)	European Conference on Computer Vision	Springer
Aneja et al. (2018)	Conference on Computer Vision and Pattern Recognition	IEEE
Bai and An (2018)	Neurocomputing	Elsevier
Banerjee and Lavie (2005)	Proceedings of the Second Workshop on Intrinsic and Extrinsic evaluation measures for Machine Translation and/or Summarization	ACL
Ben et al. (2022)	IEEE Transactions on Multimedia	IEEE
Bernardi et al. (2016)	Journal of Artificial Intelligence Research	AAAI Press
Bhosale and Patnaik (2022)	Neural Processing Letters	Springer
Bhosale and Patnaik (2022)	International Conference on Advanced Computing and Communication Systems	IEEE
Bhosale et al. (2022)	International Conference on Advanced Computing and Communication Systems	IEEE
Bin et al. (2022)	IEEE Transactions on Circuits and Systems for Video Technology	IEEE
Chen and Zitnick (2015)	Conference on Computer Vision and Pattern Recognition	IEEE
Cheng et al. (2021)	IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing	IEEE
Cornia et al. (2020)	Conference on Computer Vision and Pattern Recognition	IEEE
Dai et al. (2017)	International Conference on Computer Vision	IEEE
Devlin et al. (2015)	Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing	ACL
Dhir et al. (2019)	Computacióny Sistemas	IPN
Jeff et al. (2017)	IEEE Transactions on Pattern Analysis and Machine Intelligence	IEEE
Effendi et al. (2021)	IEEE Access	IEEE
Elliott and de Vries (2015)	International Joint Conference on Natural Language Processing	ACL
Elliott et al. (2015)	Preprint	Arxiv
Elliott and Keller (2013)	Conference on Empirical Methods in Natural Language Processing	ACL
Fang et al. (2015)	Conference on Computer Vision and Pattern Recognition	IEEE
Fang et al. (2022)	Conference on Computer Vision and Pattern Recognition	IEEE
Farhadi et al. (2010)	European Conference on Computer Vision	Springer

Table 10 (continued)

Author(s)	Journal/Conference	Publisher
Fei (2021)	AAAI Conference on Artificial Intelligence	AAAI Press
Felzenszwalb et al. (2009)	IEEE Transactions on Pattern Analysis and Machine Intelligence	IEEE
Gan et al. (2016)	Conference on Computer Vision and Pattern Recognition	IEEE
Gao et al. (2019)	AAAI Conference on Artificial Intelligence	AAAI Press
Gilberto et al. (2015)	Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies	ACL
Gong et al. (2014)	European Conference on Computer Vision	Springer
Grubinger et al. (2006)	International workshop ontoImage	AAP Publisher
Gu et al. (2017)	International Conference on Computer Vision	IEEE
Guo et al. (2020a)	International Joint Conference on Artificial Intelligence	IJCAI
Guo et al. (2020b)	Conference on Computer Vision and Pattern Recognition	IEEE
Gupta et al. (2012)	AAAI Conference on Artificial Intelligence	AAAI Press
Haque et al. (2021)	IEEE Access	IEEE
Herdade et al. (2019)	Conference and Workshop on Advances in Neural Information Processing Systems	MIT Press
Hodosh et al. (2013)	Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence	Springer
Hossain et al. (2018)	ACM Computing Surveys	ACM
Hossain et al. (2021)	IEEE Access	IEEE
Hou et al. (2021)	IEEE Access	IEEE
Hoxha et al. (2020)	IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing	IEEE
Huang et al. (2019)	International Conference on Computer Vision	IEEE
Huang et al. (2021)	IEEE Geoscience and Remote Sensing Letters	IEEE
Jia et al. (2015)	International Conference on Computer Vision	IEEE
Jiang et al. (2021)	IEEE Access	IEEE
Jin et al. (2015)	Preprint	Arxiv
Karpathy and Fei-Fei (2014)	IEEE Transactions on Pattern Analysis and Machine Intelligence	IEEE
Karpathy et al. (2014)	Conference and Workshop on Advances in Neural Information Processing Systems	MIT Press

Table 10 (continued)

Author(s)	Journal/Conference	Publisher
Kastner et al. (2021)	IEEE Access	IEEE
Kaur et al. (2021)	INFOCOMP Journal of Computer Science	University of Lavras
Ke et al. (2019)	Preprint	Arxiv
Kiros et al. (2014)	International Conference on Machine Learning	PLMR
Kiros et al. (2014)	Preprint	Arxiv
Krishna et al. (2017)	International Journal of Computer Vision	Springer
Kulkarni et al. (2013)	IEEE Transactions on Pattern Analysis and Machine Intelligence	IEEE
Kumar and Goel (2018)	International Journal of Hybrid Intelligent Systems	ACM
Kusner et al. (2015)	International Conference on Machine Learning	PLMR
Kuznetsova et al. (2012)	Annual Meeting of the Association for Computational Linguistics	ACL
Kuznetsova et al. (2014)	Transactions of the Association for Computational Linguistics	MIT Press
Lan et al. (2017)	ACM international conference on Multimedia	ACM
Lebret et al. (2015)	Preprint	Arxiv
Lebret et al. (2015)	International Conference on Machine Learning	PLMR
Li et al. (2019)	International Conference on Computer Vision	IEEE
Li et al. (2019)	Applied Sciences	MDPI
Li et al. (2019)	IEEE Transactions on Emerging Topics in Computational Intelligence	IEEE
Li et al. (2011)	Conference on Computational Natural Language Learning	ACL
Li et al. (2021)	IEEE Access	IEEE
Li et al. (2016)	International Conference on Multimedia Retrieval	ACM
Li et al. (2019)	IEEE Transactions on Multimedia	IEEE
Li et al. (2021)	IEEE Transactions on Geoscience and Remote Sensing	IEEE
Lin (2004)	Text Summarization Branches Out	ACL
Lin et al. (2015)	British Machine Vision Conference	BMVA
Lin et al. (2014)	European Conference on Computer Vision	Springer

Table 10 (continued)

Author(s)	Journal/Conference	Publisher
Liu et al. (2017a)	AAAI Conference on Artificial Intelligence	AAAI Press
Liu et al. (2019)	International Joint Conference on Artificial Intelligence	Springer
Liu et al. (2021)	IEEE Transactions on Image Processing	IEEE
Liu et al. (2017b)	International Conference on Computer Vision	IEEE
Liu et al. (2019)	The Visual Computer: International Journal of Computer Graphics	Springer
Lu et al. (2017)	Conference on Computer Vision and Pattern Recognition	IEEE
Lu et al. (2018)	Conference on Computer Vision and Pattern Recognition	IEEE
Luo et al. (2021)	AAAI Conference on Artificial Intelligence	AAAI Press
Ma et al. (2021)	IEEE Geoscience and Remote Sensing Letters	IEEE
Mao et al. (2018)	Preprint	Arxiv
Mason and Charniak (2014)	Association for Computational Linguistics	ACL
Min et al. (2021)	IEEE Access	IEEE
Mishra et al. (2021)	Computers & Electrical Engineering	Elsevier
Mitchell et al. (2012)	Conference of the European Chapter of the Association for Computational Linguistics	ACL
Miyazaki and Shimizu (2016)	Annual Meeting of the Association for Computational Linguistics	ACL
Ordonez et al. (2011)	Advances in Neural Information Processing Systems	MIT Press
Pan et al. (2020)	Conference on Computer Vision and Pattern Recognition	IEEE
Papineni et al. (2002)	Annual Meeting of the Association for Computational Linguistics	ACL
Park et al. (2017)	IEEE Conference on Computer Vision and Pattern Recognition	IEEE
Park et al. (2021)	IEEE Access	IEEE
Patterson et al. (2014)	International Journal of Computer Vision	Springer
Pedersoli et al. (2017)	Preprint	HAL
Qin et al. (2019)	Conference on Computer Vision and Pattern Recognition	IEEE
Rashtchian et al. (2010)	Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk	ACL
Rathi (2020)	International Conference on Computer, Electrical & Communication Engineering	IEEE

Table 10 (continued)

Author(s)	Journal/Conference	Publisher
Ren et al. (2017)	Conference on Computer Vision and Pattern Recognition	IEEE
Russell et al. (2008)	International Journal of Computer Vision	Springer
Sammani and Melas-Kyriazi (2020)	Conference on Computer Vision and Pattern Recognition	IEEE
Shetty et al. (2017)	International Conference on Computer Vision	IEEE
Socher et al. (2014)	Transactions of the Association for Computational Linguistics	MIT Press
Jing and Li (2021)	IEEE Access	IEEE
Sumbul et al. (2021)	IEEE Transactions on Geoscience and Remote Sensing	IEEE
Tavakoli et al. (2017)	Preprint	Arxiv
Tsutsui and Crandall (2017)	Preprint	Arxiv
Ushiku et al. (2015)	International Conference on Computer Vision	IEEE
van Miltenburg et al. (2017)	International Conference on Natural Language Generation	ACL
Vedantam et al. (2015)	IEEE Conference on Computer Vision and Pattern Recognition	IEEE
Verma and Jawahar (2014)	British Machine Vision Conference	BMVA
Vinyals et al. (2015)	Conference on Computer Vision and Pattern Recognition	IEEE
Wang et al. (2020)	IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing	IEEE
Wang and Gu (2021)	International Joint Conference on Neural Networks	IEEE
Wang et al. (2018)	ACM international conference on Multimedia	ACM
Wang et al. (2019)	Proceedings of the Beyond Vision and LANguage: inTEgrating Real-world kNowledge	ACL
Wang et al. (2020)	AAAI Conference on Artificial Intelligence	AAAI Press
Wang et al. (2021)	IEEE Transactions on Geoscience and Remote Sensing	IEEE
Wang and Chan (2018)	Preprint	Arxiv
Wang et al. (2019)	AAAI Conference on Artificial Intelligence	AAAI Press
Wang et al. (2022)	AAAI Conference on Artificial Intelligence	AAAI Press
Wang et al. (2017)	Conference on Computer Vision and Pattern Recognition	IEEE
Jie et al. (2021)	IEEE Transactions on Multimedia	IEEE

Table 10 (continued)

Author(s)	Journal/Conference	Publisher
Lingxiang et al. (2021)	IEEE Transactions on Circuits and Systems for Video Technology	IEEE
Qi et al. (2017)	IEEE Transactions on Pattern Analysis and Machine Intelligence	IEEE
Xu et al. (2015)	International Conference on Machine Learning	JMLR
Yang et al. (2021)	IEEE Transactions on Multimedia	IEEE
Yang et al. (2019a)	Conference on Computer Vision and Pattern Recognition	IEEE
Yang et al. (2019b)	International Conference on Computer Vision	IEEE
Yang et al. (2011)	Conference on Empirical Methods in Natural Language Processing	ACL
Yao et al. (2018)	European Conference on Computer Vision	Springer
Yao et al. (2019)	IEEE/CVF International Conference on Computer Vision	IEEE
Yatskar et al. (2014)	Joint Conference on Lexical and Computational Semantics	ACL
Ye et al. (2018)	IEEE Transactions on Image Processing	IEEE
Yoshikawa et al. (2017)	Annual Meeting of the Association for Computational Linguistics	ACL
Jun et al. (2020)	IEEE Transactions on Circuits and Systems for Video Technology	IEEE
Yuan et al. (2020)	IEEE Access	IEEE
Yumeng et al. (2021)	IEEE Access	IEEE
Zhang et al. (2021)	IEEE Transactions on Multimedia	IEEE
Zhang et al. (2021)	Conference on Computer Vision and Pattern Recognition	IEEE
Zhao et al. (2021)	IEEE Transactions on Image Processing	IEEE
Zhou et al. (2020)	Conference on Computer Vision and Pattern Recognition	IEEE
Zhou et al. (2021)	IEEE Access	IEEE
Zohourianshahzadi and Kalita (2021)	Artificial Intelligence Review	Springer

Table 11 A frequency distribution table of the article publishers cited in this survey

S. no.	Publisher	URL	Frequency
1	IEEE	https://www.ieee.org/	76
2	ACL	https://www.aclweb.org/portal/	18
3	Springer	https://link.springer.com/	13
4	AAAI Press	https://www.aaai.org/Press/press.php	10
5	Arxiv	https://arxiv.org/	6
6	ACM	https://www.acm.org/	5
7	MIT Press	https://mitpress.mit.edu/	5
8	MLR Press	https://proceedings.mlr.press/	4
9	Elsevier	https://www.elsevier.com	3
10	BMVA	https://britishmachinevisionassociation.github.io/	2
11	AAP Publication	https://publishers.org/	1
12	IPN	https://www.ipn.mx/	1
13	MDPI	https://www.mdpi.com/	1
14	University of Lavras	https://ufla.br/en/ufla/	1
15	IJCAI	https://www.ijcai.org/	1
Total			147

Table 12 A frequency distribution table of article types cited in this survey

S. no.	Article type	Frequency
1	Journal	58
2	Conference	83
3	Preprint	6
Total		147

Acknowledgements The authors extend sincere gratitude to the Editor and Reviewers for their insightful remarks and helpful opinions, which contributed to the enhancement of the work.

Declarations

Conflict of interest All the authors declare that they do not have any conflict of interest.

References

Alam M, Samad MD, Vidyaratne L, Glandon A, Iftekharuddin KM (2020) Survey on deep neural networks in speech and vision systems. Neurocomputing 417:302–321

Amirian S, Rasheed K, Taha TR, Arabnia HR (2020) Automatic image and video caption generation with deep learning: a concise review and algorithmic overlap. IEEE Access 8:218386–218400

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

Aneja J, Deshpande A, Schwing AG (2018) Convolutional image captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5561–5570

Bai S, An S (2018) A survey on automatic image caption generation. Neurocomputing 311:291–304

Banerjee S, Lavie A (2005) Meteor: an automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, pp 65–72
Ben H, Pan Y, Li Y, Yao T, Hong R, Wang M, Mei T (2022) Unpaired image captioning with semantic-constrained self-learning. IEEE Trans Multimedia 24:904–916
Bernardi R, Cakici R, Elliott D, Erdem A, Erdem E, Ikizler-Cinbis N, Keller F, Muscat A, Plank B (2016) Automatic description generation from images: a survey of models, datasets, and evaluation measures. J Artif Intell Res 55:409–442
Bhosale YH, Patnaik KS (2022) Application of deep learning techniques in diagnosis of covid-19 (coronavirus): a systematic review. Neural Process Lett 2:1–53
Bhosale YH, Patnaik KS (2022) Iot deployable lightweight deep learning application for covid-19 detection with lung diseases using raspberrypi. In: 2022 International Conference on IoT and Blockchain Technology (ICIBT), pp 1–6
Bhosale YH, Zanwar S, Ahmed Z, Nakrani M, Bhuyar D, Shinde U (2022) Deep convolutional neural network based covid-19 classification from radiology x-ray images for IOT enabled devices. In: 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS), vol 1, pp 1398–1402
Bin Y, Ding Y, Peng B, Peng L, Yang Y, Chua T-S (2022) Entity slot filling for visual captioning. IEEE Trans Circuits Syst Video Technol 32(1):52–62
Bryan CR, Antonio T, Murphy Kevin P, Freeman William T (2008) Labelme: a database and web-based tool for image annotation. Int J Comput Vision 77(1–3):157–173
Chen X, Zitnick C.L (2015) Mind's eye: a recurrent visual representation for image caption generation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 2422–2431, Boston. IEEE
Cheng Q, Zhou Y, Peng F, Yuan X, Zhang L (2021) A deep semantic alignment network for the cross-modal image-text retrieval in remote sensing. IEEE J Select Top Appl Earth Observations Remote Sens 14:4284–4297
Cornia M, Stefanini M, Baraldi L, Cucchiara R (2020) Meshed-memory transformer for image captioning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10578–10587
Dai B, Fidler S, Urtasun R, Lin D (2017) Towards diverse and natural image descriptions via a conditional GAN. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp 2989–2998, Venice. IEEE
Devlin J, Cheng H, Fang H, Gupta S, Deng L, He X, Zweig G, Mitchell M (2015) Language models for image captioning: the quirks and what works. In: Annual Meeting of the Association for Computational Linguistics
Dhir R, Mishra SK, Saha S, Bhattacharyya P (2019) A deep attention based framework for image caption generation in Hindi language. Computación y Sistemas 23(3):125
Effendi J, Sakti S, Nakamura S (2021) End-to-end image-to-speech generation for untranscribed unknown languages. IEEE Access 55(9):55144–55154
Elliott D, de Vries A (2015) Describing images using inferred visual dependency representations. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (vol 1: Long Papers), pp 42–52
Elliott D, Frank S, Hasler E (2015) Multilingual image description with neural sequence models. *arXiv preprint*arXiv:1510.04709
Elliott D, Keller F (2013) Image description using visual dependency representations. In: Proceedings of the 2013 conference on empirical methods in natural language processing, pp 1292–1302
Fang H, Gupta S, Iandola F, Srivastava RK, Deng L, Dollár P, Gao J, He X, Mitchell M, Platt JC, et al (2015) From captions to visual concepts and back. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1473–1482
Fang Z, Wang J, Hu X, Liang L, Gan Z, Wang L, Yang Y, Liu Z (2022) Injecting semantic concepts into end-to-end image captioning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 18009–18019
Farhadi A, Hejrati M, Sadeghi MA, Young P, Rashtchian C, Hockenmaier J, Forsyth DA (2010) Every picture tells a story: generating sentences from images. In: European Conference on Computer Vision
Fei Z (2021) Partially non-autoregressive image captioning. Proc AAAI Conf Artif Intell 35:1309–1316
Felzenszwalb PF, Girshick RB, McAllester D, Ramanan D (2009) Object detection with discriminatively trained part-based models. IEEE Trans Pattern Anal Mach Intell 32(9):1627–1645

Gan Z, Gan C, He X, Pu Y, Tran K, Gao J, Carin L, Deng L (2016) Semantic compositional networks for visual captioning. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp 1141–1150

Gao L, Fan K, Song J, Liu X, Xu X, Shen HT (2019) Deliberate attention networks for image captioning. In: AAAI Conference on Artificial Intelligence

Gilberto L, Ortiz M, Wolff C, Lapata M (2015) Learning to interpret and describe abstract scenes. In: Proceedings of the 2015 conference of the north american chapter of the association for computational linguistics: human language Technologies, pps 1505–1515, Denver, Colorado, 2015. Association for Computational Linguistics

Girish K, Visruth P, Vicente O, Sagnik D, Siming L, Yejin C, Berg Alexander C, Berg Tamara L (2013) Babytalk: understanding and generating simple image descriptions. IEEE Trans Pattern Anal Mach Intell 35(12):2891–2903

Gong Y, Wang L, Hodosh M, Hockenmaier Julia, Lazebnik Svetlana (2014) Improving image-sentence embeddings using large weakly annotated photo collections. In: European conference on computer vision, pp 529–545. Springer

Grubinger M, Clough PM, Deselaers T (2006) The iapr tc-12 benchmark: a new evaluation resource for visual information systems. In: International workshop ontoImage, volume 2

Gu J, Wang G, Cai J, Chen T (2017) An empirical study of language cnn for image captioning. In: 2017 IEEE international conference on computer vision (ICCV), pp 1231–1240, Venice. IEEE

Guo L, Liu J, Zhu X, He X, Jiang J, Lu H (2021) Non-autoregressive image captioning with counterfactuals-critical multi-agent learning. In: Proceedings of the twenty-ninth international joint conference on artificial intelligence, IJCAI'20, 2021

Guo L, Liu J, Zhu X, Yao P, Lu S, Lu H (2020) Normalized and geometry-aware self-attention network for image captioning. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 10324–10333, Seattle. IEEE

Gupta A, Verma Y, Jawahar CV (2012) Choosing linguistics over vision to describe images. In: Proceedings of the AAAI conference on artificial intelligence

Haque AUI, Ghani S, Saeed M (2021) Image captioning with positional and geometrical semantics. IEEE Access 9:160917–160925

Herdade S, Kappeler A, Boakye K, Soares J (2019) Image captioning: Transforming objects into words. In: Conference and workshop on neural information processing systems

Hodosh M, Young P, Hockenmaier J (2013) Framing image description as a ranking task: data, models and evaluation metrics. J Artif Intell Res 47:853–899

Hossain MZ, Sohel F, Shiratuddin MF, Laga H (2019) A comprehensive survey of deep learning for image captioning. ACM Comput Surveys (CSUR) 51:1–36

Hossain MZ, Sohel F, Shiratuddin MF, Laga H, Bennamoun M (2021) Text to image synthesis for improved image captioning. IEEE Access 9:64918–64928

Hou D, Zhao Z, Liu Y, Chang F, Sanyuan H (2021) Automatic report generation for chest X-ray images via adversarial reinforcement learning. IEEE Access 9:21236–21250

Hoxha G, Melgani F, Demir B (2020) Toward remote sensing image retrieval under a deep image captioning perspective. IEEE J Select Top Appl Earth Observ Remote Sens 13:4462–4475

Huang Wei, Wang Qi, Li X (2021) Denoising-based multiscale feature fusion for remote sensing image captioning. IEEE Geosci Remote Sens Lett 18(3):436–440

Huang L, Wang W, Chen J, Wei X-Y (2019) Attention on attention for image captioning. In: 2019 IEEE/cvf international conference on computer vision (ICCV), pp 4633–4642

Jeff D, Anne HL, Marcus R, Subhashini V, Sergio G, Kate S, Trevor D (2017) Long-term recurrent convolutional networks for visual recognition and description. IEEE Trans Pattern Anal Mach Intell 39(4):677–691

Jia X, Gavves E, Fernando B, Tuytelaars T (2015) Guiding the long-short term memory model for image caption generation. In: 2015 IEEE international conference on computer vision (ICCV)

Jiang W, Li X, Haifeng H, Qiang L, Liu B (2021) Multi-gate attention network for image captioning. IEEE Access 9:69700–69709

Jie W, Chen T, Hefeng W, Yang Z, Luo G, Lin Liang (2021) Fine-grained image captioning with global-local discriminative objective. IEEE Trans Multimedia 23:2413–2427

Jin J, Fu K, Cui R, Sha F, Zhang C (2015) Aligning where to see and what to tell: image caption with region-based attention and scene factorization. arXiv:1506.06272 [cs, stat], June 2015

Jing Su, Li Jing (2021) Show auto-adaptive and tell: learned from the SEM image challenge. IEEE Access 9:51494–51500

Jun Y, Li J, Zhou Y, Huang Q (2020) Multimodal transformer with multi-view visual representation for image captioning. IEEE Trans Circuits Syst Video Technol 30(12):4467–4480

Karpathy A, Fei-Fei L (2014) Deep visual-semantic alignments for generating image descriptions. IEEE Trans Pattern Anal Mach Intell 39:664–676
Karpathy A, Joulin A, Fei-Fei LF (2014) Deep fragment embeddings for bidirectional image sentence mapping. Adv Neural Inform Process Syst 27:1
Kastner MA, Umemura K, Ide I, Kawanishi Y, Hirayama T, Doman Keisuke, Deguchi Daisuke, Murase Hiroshi, Satoh Shin'Ichi (2021) Imageability- and length-controllable image captioning. IEEE Access 9:162951–162961
Kaur M, Josan G, Kaur J (2021) Automatic Punjabi caption generation for sports images. INFOCOMP J Comput Sci 20(1):2
Ke L, Pei W, Li R, Shen X, Tai Y-W (2019) Reflective decoding network for image captioning. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 8888–8897
Kiros R, Salakhutdinov R, Zemel RS (2014) Multimodal neural language models. In: International Conference on Machine Learning
Kiros R, Salakhutdinov R, Zemel RS (2014) Unifying visual-semantic embeddings with multimodal neural language models. arXiv:1411.2539 [cs], November 2014
Kumar A, Goel S (2018) A survey of evolution of image captioning techniques. Int J Hybrid Intell Syst 14(3):123–139
Kusner M, Sun Y, Kolkin N, Weinberger K (2015) From word embeddings to document distances. In: International conference on machine learning, pp 957–966. PMLR
Kuznetsova P, Ordonez V, Berg TL, Choi Y (2014) Treetalk : composition and compression of trees for image descriptions. Trans Assoc Comput Linguist 2:351–362
Kuznetsova P, Ordonez V, Berg AC, Berg TL, Choi Y (2012) Collective generation of natural image descriptions. In: Annual meeting of the association for computational linguistics
Lan W, Li X, Dong J (2017) Fluency-guided cross-lingual image captioning. In: Proceedings of the 25th ACM international conference on multimedia, pp 1549–1557
Lebret R, Pinheiro PHO, Collobert R (2015) Phrase-based image captioning. In: International conference on machine learning
Lebret R, Pinheiro PO, Collobert R (2015) Simple image description generator via a linear phrase-based approach. arXiv:1412.8419 [cs], April
Li X, Chaoxi X, Wang X, Lan W, Jia Z, Yang G, Jieping X (2019) COCO-CN for cross-lingual image tagging, captioning, and retrieval. IEEE Trans Multimedia 21(9):2347–2360
Li S, Tao Z, Li K, Yun F (2019) Visual to text: survey of image and video captioning. IEEE Trans Emerging Top Comput Intell 3(4):297–312
Li J, Yao P, Guo L, Zhang W (2019) Boosted transformer for image captioning. Appl Sci 9(16):3260
Li X, Zhang X, Huang W, Wang Q (2021) Truncation cross entropy loss for remote sensing image captioning. IEEE Trans Geosci Remote Sens 59(6):5246–5257
Li W, Zhaowei Q, Song H, Wang P, Xue B (2021) The traffic scene understanding and prediction based on image captioning. IEEE Access 9:1420–1427
Li G, Zhu L, Liu P, Yang Y (2019) Entangled transformer for image captioning. In: 2019 IEEE/CVF international conference on computer vision (ICCV), pp 8927–8936
Li S, Kulkarni G, Berg T, Berg A, Choi Y (2011) Composing simple image descriptions using web-scale n-grams. In: Proceedings of the fifteenth conference on computational natural language learning, pp 220–228
Li X, Lan W, Dong J, Liu H (2016) Adding Chinese captions to images. In: Proceedings of the 2016 ACM on international conference on multimedia retrieval, pp 271–275
Lin C-Y (2004) Rouge: a package for automatic evaluation of summaries. In: Text summarization branches out, pp 74–81
Lin D, Fidler S, Kong C, Urtasun R (2015) Generating multi-sentence natural language descriptions of indoor scenes. In: Proceedings of the British machine vision conference (BMVC), pp 93.1–93.13. BMVA Press, September 2015
Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: Common objects in context. In: European conference on computer vision, pp 740–755. Springer
Lingxiang W, Min X, Sang L, Yao T, Mei T (2021) Noise augmented double-stream graph convolutional networks for image captioning. IEEE Trans Circuits Syst Video Technol 31(8):3118–3127
Liu X, Qingyang X, Wang N (2019) A survey on deep neural network-based image captioning. Visual Comput 35(3):445–470
Liu H, Zhang S, Lin K, Wen J, Li J, Xiaolin H (2021) Vocabulary-wide credit assignment for training image captioning models. IEEE Trans Image Process 30:2450–2460
Liu C, Mao J, Sha F, Yuille A (2017) Attention correctness in neural image captioning. In: Thirty-first AAAI conference on artificial intelligence,

Liu F, Ren X, Liu Y, Lei K, Sun X (2019) Exploring and distilling cross-modal information for image captioning. In: International joint conference on artificial intelligence

Liu S, Zhu Z, Ye N, Guadarrama S, Murphy K (2017) Improved image captioning via policy gradient optimization of SPIDEr. In: 2017 IEEE international conference on computer vision (ICCV), pp 873–881, Venice. IEEE

Lu J, Xiong C, Parikh D, Socher R (2017) Knowing when to look: adaptive attention via a visual sentinel for image captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 375–383

Lu J, Yang J, Batra D, Parikh D (2018) Neural baby talk. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7219–7228

Luo Y, Ji J, Sun X, Cao L, Yongjian W, Huang F, Lin CW, Ji R (2021) Dual-level collaborative transformer for image captioning. Proc AAAI Conf Artif Intell 35:2286–2293

Ma X, Zhao R, Shi Z (2021) Multiscale methods for optical remote-sensing image captioning. IEEE Geosci Remote Sens Lett 18(11):2001–2005

Mao J, Xu W, Yang Y, Wang J, Huang Z, Yuille A (2018) Deep captioning with multimodal recurrent neural networks (m-RNN). In: Proceedings of the international conference on learning representations

Mason R, Charniak E (2014) Nonparametric method for data-driven image captioning. In: Proceedings of the 52nd annual meeting of the association for computational linguistics (Volume 2: Short Papers), pp 592–598, Baltimore, Maryland. Association for Computational Linguistics

Min K, Dang M, Moon H (2021) Deep learning-based short story generation for an image using the encoder-decoder structure. IEEE Access 9:113550–113557

Mishra SK, Dhir R, Saha S, Bhattacharyya P, Singh AK (2021) Image captioning in Hindi language using transformer networks. Comput Electr Eng 92:107114

Mitchell M, Dodge J, Goyal A, Yamaguchi K, Stratos K, Han X, Mensch A, Berg A, Berg T, Daumé IIIH (2012) Midge: generating image descriptions from computer vision detections. In: Proceedings of the 13th conference of the European chapter of the association for computational linguistics, pp 747–756

Miyazaki T, Shimizu N (2016) Cross-lingual image caption generation. In: Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers), pp 1780–1790

Ordonez V, Kulkarni G, Berg T (2011) Im2text: describing images using 1 million captioned photographs. Advn Neural Inform Process Syst 56:25

Pan Y, Yao T, Li Y, Mei T, (2020) X-linear attention networks for image captioning. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 10968–10977, 2020

Papineni K, Roukos S, Ward T, Zhu W-Ji(2002) Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the association for computational linguistics, pp 311–318

Park H, Kim K, Park S, Choi J (2021) Medical image captioning model to convey more details: methodological comparison of feature difference generation. IEEE Access 9:150560–150568

Park CC, Kim B, Kim G (2017) Attend to you: personalized image captioning with context sequence memory networks. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp 6432–6440

Patterson G, Chen X, Hang S, Hays J (2014) The SUN attribute database: beyond categories for deeper scene understanding. Int J Comput Vision 108(1–2):59–81

Pedersoli M, Lucas T, Schmid C, Verbeek J (2017) Areas of attention for image captioning. In: Proceedings of the IEEE international conference on computer vision, pp 1242–1250

Qi W, Shen C, Wang P, Dick A, Van Den Hengel A (2017) Image captioning and visual question answering based on attributes and external knowledge. IEEE Trans Pattern Anal Mach Intell 40(6):1367–1381

Qin Y, Du J, Zhang Y, Lu H (2019) Look back and predict forward in image captioning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8367–8375

Ranjay K, Yuke Z, Oliver G, Justin J, Kenji H, Joshua K, Stephanie C, Yannis K, Li-Jia L, Shamma DA et al (2017) Visual genome: connecting language and vision using crowdsourced dense image annotations. Int J Comput Vision 123(1):32–73

Rashtchian C, Young P, Hodosh M, Hockenmaier J (2010) Collecting image annotations using amazon's mechanical turk. In: Proceedings of the NAACL HLT 2010 Workshop on creating speech and language data with amazon's mechanical Turk, pp 139–147, 2010

Rathi A (2020) Deep learning apporach for image captioning in Hindi language. In: 2020 international conference on computer, electrical communication engineering (ICCECE), pp 1–8, 2020

Ren Z, Wang X, Zhang N, Lv X, Li L-J (2017) Deep reinforcement learning-based image captioning with embedding reward. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp 1151–1159, Honolulu. IEEE

Sammani F, Melas-Kyriazi L (2020) Show, edit and tell: a framework for editing image captions. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 4807–4815, Seattle. IEEE

Shetty R, Rohrbach M, Hendricks LA, Fritz M, Schiele B (2017) Speaking the same language: matching machine to human captions by adversarial training. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp 4155–4164, Venice 2017. IEEE

Socher R, Karpathy A, Le QV, Manning CD, Andrew YN (2014) Grounded compositional semantics for finding and describing images with sentences. Trans Assoc Comput Linguist 2:207–218

Sumbul G, Nayak S, Demir B (2021) SD-RSIC: summarization-driven deep remote sensing image captioning. IEEE Trans Geosci Remote Sens 59(8):6922–6934

Tavakoli HR, Shetty R, Borji A, Laaksonen J (2017) Paying attention to descriptions generated by image captioning models. In: Proceedings of the IEEE international conference on computer vision, pp 2487–2496

Tsutsui S, Crandall D (2017) Using artificial tokens to control languages for multilingual image caption generation. *arXiv preprint*arXiv:1706.06275*, 2017*

Ushiku Y, Yamaguchi M, Mukuta Y, Harada T (2015) Common subspace for model and similarity: phrase learning for caption generation from images. In: 2015 IEEE international conference on computer vision (ICCV), pp 2668–2676, Santiago, Chile 2015. IEEE

van Miltenburg E, Elliott D, Vossen P (2017) Cross-linguistic differences and similarities in image descriptions. In: International conference on natural language generation, 2017

Vedantam R, Zitnick CL, Parikh D (2015) Cider: consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4566–4575, 2015

Verma Y, Jawahar CV (2014) Im2Text and Text2Im: associating images and texts for cross-modal retrieval. In Proceedings of the British machine vision conference 2014, pp 97.1–97.13, Nottingham, 2014. British Machine Vision Association

Vinyals O, Toshev A, Bengio S, Erhan D (2015) Show and tell: a neural image caption generator. In: 2015 IEEE conference on computer vision and pattern recognition (CVPR), pp 3156–3164, Boston. IEEE

Wang L, Bai Z, Zhang Y, Hongtao L (2020) Show, recall, and tell: image captioning with recall mechanism. Proc AAAI Conf Artif Intell 34(07):12176–12183

Wang W, Chen Z, Haifeng H (2019) Hierarchical attention network for image captioning. Proc AAAI Conf Artif Intell 33:8957–8964

Wang Q, Huang W, Zhang X, Li X (2021) Word-sentence framework for remote sensing image captioning. IEEE Trans Geosci Remote Sens 59(12):10532–10543

Wang Y, Jungang X, Sun Y (2022) End-to-end transformer based model for image captioning. Proc AAAI Conf Artif Intell 36:2585–2594

Wang B, Zheng X, Bo Q, Xiaoqiang L (2020) Retrieval topic recurrent memory network for remote sensing image captioning. IEEE J Select Top Appl Earth Observ Remote Sens 13:256–270

Wang C, Gu X (2021) An image captioning approach using dynamical attention. In: 2021 international joint conference on neural networks (IJCNN), pp 1–8, 2021

Wang C, Yang H, Meinel C (2018) Image captioning with deep bidirectional lstms and multi-task learning. In: ACM transactions on multimedia computing, communications, and applications (TOMM), 14(2s):1–20

Wang D, Beck D, Cohn T (2019) On the role of scene graphs in image captioning. In: Proceedings of the beyond vision and language: integrating real-world knowledge (LANTERN)

Wang Q, Chan AB (2018) CNN+CNN: convolutional decoders for image captioning. arXiv:1805.09019 [cs], May

Wang Y, Lin Z, Shen X, Cohen S, Cottrell GW (2017) Skeleton key: image captioning by skeleton-attribute decomposition. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp 7378–7387, Honolulu. IEEE

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

Yang L, Wang H, Tang P, Li Qinyu (2021) CaptionNet: a tailor-made recurrent neural network for generating image descriptions. IEEE Trans Multimedia 23:835–845

Yang X, Tang K, Zhang H, Cai J (2019) Auto-encoding scene graphs for image captioning. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 10677–10686, Long Beach 2019. IEEE

Yang X, Zhang H, Cai J (2019) Learning to collocate neural modules for image captioning. In: 2019 IEEE/CVF international conference on computer vision (ICCV), pp 4249–4259, Seoul 2019. IEEE

Yang Y, Teo C, Daumé H, Aloimonos Y (2011) Corpus-guided sentence generation of natural images. In: Proceedings of the 2011 conference on empirical methods in natural language processing, pp 444–454
Yao T, Pan Y, Li Y, Mei T (2018) Exploring visual relationship for image captioning. In: Proceedings of the European conference on computer vision (ECCV), pp 684–699, 2018
Yao T, Pan Y, Li Y, Mei T (2019) Hierarchy parsing for image captioning. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 2621–2629
Yatskar M, Galley M, Vanderwende L, Zettlemoyer L (2014) See no evil, say no evil: Description generation from densely labeled images. In : Proceedings of the Third Joint Conference on Lexical and Computational Semantics (* SEM 2014), pp 110–120
Ye S, Han J, Liu N (2018) Attentive linear transformation for image captioning. IEEE Trans Image Process 27(11):5514–5524
Yoshikawa Y, Shigeto Y, Takeuchi A (2017) STAIR captions: constructing a large-scale Japanese image caption dataset. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp 417–421, Vancouver. Association for Computational Linguistics
Yuan Z, Li X, Wang Q (2020) Exploring multi-level attention and semantic relationship for remote sensing image captioning. IEEE Access 8:2608–2620
Yumeng Z, Jing Y, Shuo G, Limin L (2021) News image-text matching with news knowledge graph. IEEE Access 9:108017–108027
Zhang J, Mei K, Zheng Y, Fan J (2021) Integrating part of speech guidance for image captioning. IEEE Trans Multimedia 23:92–104
Zhang X, Sun X, Luo Y, Ji J, Zhou Y, Wu Y, Huang F, Ji R (2021) Rstnet: Captioning with adaptive attention on visual and non-visual words. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 15465–15474
Zhao W, Xinxiao W, Luo J (2021) Cross-domain image captioning via cross-modal retrieval and model adaptation. IEEE Trans Image Process 30:1180–1192
Zhou Z, Liang X, Wang C, Xie W, Wang S, Ge S, Zhang Y (2021) An image captioning model based on bidirectional depth residuals and its application. IEEE Access 9:25360–25370
Zhou Y, Wang M, Liu D, Hu Z, Zhang H (2020) More grounded image captioning by distilling image-text matching model. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)
Zohourianshahzadi Z, Kalita JK (2021) Neural attention for image captioning: review of outstanding methods. Artif Intell Rev 55(5):3833–3862