International Journal of Computer Engineering in Research Trends
ISSN: 2349-7084/ https://doi.org/10.22362/ijcert.v10i12.912
Volume 10, Issue 12, December 2023

Research Paper

Capturing the Essence: An In-Depth Exploration of Automatic Image Captioning Techniques and Advancements

Sushma Jaiswal[1*], Harikumar Pallthadka[2], Rajesh P. Chinchewadi[3], Tarun Jaiswal[4]

[1] *Guru Ghasidas Central University, Bilaspur (C.G.) and Post-Doctoral Research Fellow, Manipur International University, Imphal, Manipur, jaiswal1302@gmail.com, https://orcid.org/0000-0002-6253-7327*
[2] *Manipur International University, Imphal, Manipur, vc@miu.edu.in, https://orcid.org/0000-0002-0705-9035.*
[3] *Manipur International University, Imphal, Manipur, rajesh.cto@miu.edu.in.*
[4] *National Institute of Technology, Raipur, tjaiswal_1207@yahoo.com, https://orcid.org/0000-0003-3963-4548*

**Sushma Jaiswal: jaiswal1302@gmail.com*

Received: 12/10/2023, Revised: 05/11/2023, Accepted: 19/12/2023 Published: 30/12/2023

Abstract: - *This comprehensive overview offers a deep dive into the quickly developing topic of automatic image captioning. It provides a thorough examination of the many approaches, datasets, and assessment criteria applied in the production of written descriptions from visual content. With an emphasis on the extensively used MS COCO dataset, the article explores the critical significance of datasets in addition to exploring the nuances of cutting-edge image captioning algorithms. It also goes over the fundamental function of assessment metrics, stressing the importance of BLEU, METEOR, ROUGE, CIDEr, and SPICE metrics in determining the caliber of created captions. This study is an invaluable tool for scholars and practitioners looking to improve the field because it offers a thorough analysis of the developments and difficulties in this field. To be more precise, we start by quickly going over the earlier traditional works using the retrieval and template. After that, research on deep learning (DL)-based image captioning is concentrated. For a thorough overview, these studies are divided into three categories: the encoder-decoder architecture, the attention techniques, and training techniques based on framework structures and training methods. The publicly accessible datasets, evaluation metrics, and those suggested for particular requirements are then summarized, and the state-of-the-art techniques are then contrasted using the MS COCO dataset, we offer a few talks about unresolved issues and potential avenues for future research.*

Keywords- *Artificial-intelligence, attention-mechanism, encoder-decoder framework, image captioning, multi-modal understanding, training strategies.*

---

1. Introduction

One of the main objectives in the fields of computer vision and natural language processing has always been to establish a connection between the written descriptions and the visual reality. The challenge of automatically producing meaningful and contextually appropriate written descriptions from visual content, known as automatic image captioning, has advanced significantly in the last few years. This thorough analysis objectives to provide a Thorough analysis of this dynamic and changing environment. Converting images to text is an extremely important procedure in several fields, such as content indexing, accessibility, and human-computer interaction. Furthermore, it has been widely used in fields including helping the blind, improving image retrieval systems, and expanding the storytelling capabilities of AI-powered virtual assistants. We take a deep dive into the approaches, datasets, and assessment criteria that characterize the SOTA in automatic image captioning in this paper. We explore the architectural underpinnings, covering the introduction of recurrent neural networks (RNNs), transformer-based models, and convolutional-neural networks (CNNs), which have completely changed the image captioning industry. We also highlight the critical role that datasets play in both training and evaluation. The Microsoft Common-Objects in Context (MS COCO) dataset is a shining example of one of these, including a wide range of images with accompanying human-generated captions. This dataset continues to drive

research in this area and has established itself as a benchmark for comparing image captioning techniques. There is a growing body of recognized evaluation metrics that are used to guarantee the quality of generated captions. We examine the fundamental metrics—BLEU, METEOR, ROUGE, CIDEr, and SPICE, among others—and talk about how important they are for determining how true captions are to human references. These metrics offer a common framework for assessing how well image captioning models function and contrasting the outcomes. We hope to provide a full grasp of the developments and difficulties that define automatic image captioning as we work through this extensive review. Researchers, developers, and enthusiasts looking to improve automatic image captioning systems will find great value in the synthesis of approaches, datasets, and assessment measures as well as insights into the current status of the field. The image captioning technical diagram shows the intricate procedures of turning visual content into textual descriptions. This incorporates CNNs for image feature extraction, RNNs or Transformer-based architectures for caption generation, and attention methods to align visual and textual information. To improve captions, reinforcement learning could be used. The overall practical diagram is displayed in Figure 1.

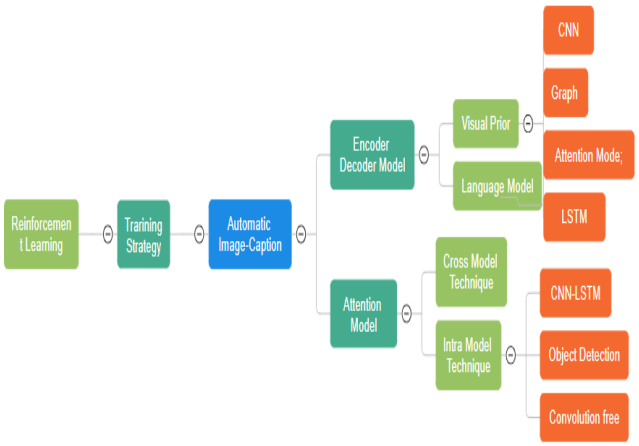

Figure 1: Overall DL-based image captioning architectural diagram

2. Literature Review

The authors [1] proposed a neural network model that selectively focuses on visual regions to generate meaningful captions. This method is based on the hypothesis that humans describe images by focusing on specific parts. Attention mechanisms help the model focus on relevant visual regions, resulting in more contextually correct and informative captions, according to the report. The model may "show" its comprehension of the visual material and "tell" a coherent tale in the generated captions by focusing on various portions of the image at each phase. The "Show, attend and tell" model presented in the paper is a major advance in image captioning because it uses CNNs for image feature extraction and RNNs for caption generation, along with attention mechanisms to improve caption quality.

The paper [2] discussed competition successes and problems, including image captioning methodologies and techniques. The MS COCO dataset is crucial to measuring and improving image captioning technology. This study helps explain image captioning's state and evolution throughout the 2015 MS COCO challenge. Computer vision and natural language processing researchers and practitioners benefit from it.

Lu et al. [3] used a visual sentinel mechanism to study adaptive attention in image captioning. The visual sentinel helps the model priorities image regions during captioning, resulting in more contextually relevant and informative captions. Improved image captioning model quality and efficiency make the paper relevant. It shows how to optimize attention processes for image captioning. The findings in this work have inspired computer vision and image captioning research.

Rennie et al. [4] used self-critical sequence training to optimize image captioning models based on their own captions. The paper examines how self-critical training can significantly enhance image captions. This paper has greatly impacted computer vision and image captioning by revealing how to train models more effectively. This self-critical sequence training method is essential to improving image captioning systems.

The Liu et al. [5] suggested the CPTR paradigm, which uses Transformer architecture for image feature extraction and caption generation. This breakthrough in image captioning distinguishes it from the typical use of CNNs and RNNs. The research shows how a complete Transformer network may better capture long-range dependencies and contextual information for image captioning. The study shows how cutting-edge DL systems may improve image captions.

In [6] studied how visual associations can improve image caption quality and context. Visual relationships between objects and elements in images are used to improve captions' descriptive effectiveness. This paper emphasises the necessity of capturing both individual items and their interactions, which could advance image captioning. This research enriches visual content descriptions, making it useful for computer vision and natural language processing researchers.

In paper [7] surveyed the crucial problem of harmonising caption words with image visuals. The research seeks to improve image caption accuracy and contextual relevance by aligning linguistic descriptions and visual information. This paper's findings can improve image captions and understanding textual-visual links. This research improves multimedia and image captioning by aligning verbal descriptions with visual content.

Yu et al. [8] used a multimodal-transformer architecture with multi-view visual representations. Authors combine visual sources to improve image caption quality and contextually. This paper shows that multimodal models and multi-view representations improve descriptive image captions, which has major implications for image captioning. This research improves computer vision and natural language processing researchers' grasp of how to integrate visual information into image captioning.

This research revisits and improves vision-language model visual representations to improve visual-textual communication [9]. The VINVL paradigm improves vision-language interactions' quality and contextually. The study shares insights about optimizing and integrating visual representations into vision-language models, adding to computer vision and natural language processing understanding. This discovery is important because vision-language models are increasingly used in image captioning and other applications [9].

Zhang et al. [10] examined the Rstnet model, which uses adaptive attention to improve captioning. Note that the attention technique is used to both visual and non-visual words, making caption production more thorough and contextually relevant.

This study investigated auto-encoded scene graphs in image captioning. The authors hope scene graphs help readers understand visual relationships and aspects, making captions more contextual and helpful [11].

Yang et al. [12] suggested collocate neural modules to improve image captioning. The authors want to create more contextually appropriate and useful image descriptions by ordering and coordinating brain modules. The paper emphasizes brain module organization and synchronization in caption production, providing new insights into image captioning algorithms. This research advances neural module optimization for image captioning, helping computer vision and natural language processing researchers and practitioners.

In [13] proposed "stack-captioning," a coarse-to-fine learning approach to improve image captioning. By sequentially refining caption production, the authors hope to provide more contextually accurate and informative image captions.

This investigation examined global-local attention, which improves image captioning. The authors use this attention mechanism to capture global and local information in photos for more contextually relevant and interesting subtitles [14].

This research improved image captioning with saliency and context attention. The authors intend to write more contextually relevant and informative image captions by focusing on salient regions and contextual information [15]. Gao et al. [16] examined "deliberate attention networks" and how they could improve image captioning. The authors used conscious attention techniques to create more contextually relevant and informative image captions. The article highlighted the relevance of conscious attention networks in refining image captions, providing useful insights. This research helped computer-vision (C V) and natural-language processing (NLP) researchers and practitioners understand how attention mechanisms could be actively integrated into image captioning. Zhang et al. [17] studied visual relationship attention to capture implicit area relationships in images to better descriptive image captions. The article helped explain how attention mechanisms might improve image caption quality and context.

X-linear attention networks [18] were used to improve image captioning in this study. These networks improved attention to provide contextually appropriate and informative image captions. By emphasizing the relevance of X-linear attention networks in increasing caption quality, the article shed light on image captioning. Herdade et al. [19] examined how captioning turns image objects into descriptive words. The research explained how visual content is translated into text. The paper, presented at the Advances in Neural Information Processing Systems conference, illuminated the process of turning visual elements into language and provided valuable insights for computer vision and natural language processing researchers and practitioners. Guo et al [20] frazzled the necessity of normalized and geometry-aware self-attention networks in image captioning. To generate more contextually relevant and meaningful image captions, these networks included geometric information and normalization into the attention mechanism.

This investigation examined the "Entangled Transformer" and image captioning. The Entangled Transformer architecture was used to provide more contextually relevant and informative image captions [21].

Cornia et al.[22] examined image captioning with the "Meshed-Memory Transformer". This method used the Meshed-Memory Transformer architecture to create more contextually relevant and informative image captions. By emphasizing the Meshed-Memory Transformer's role in improving captions, the paper shed light on image captioning. This research helped computer vision and natural language processing researchers and practitioners comprehend how new transformer designs could be used in image captioning.

The Oscar approach aligns objects and semantics in pre-training to improve vision-language tasks, according to this study. This matching of object and semantic information improves visual-text communication [23].

In [24], the authors recommended "Look Back and Predict Forward" to improve image captioning. The authors hope this method will produce more contextually relevant and useful image captions. The study demonstrated how "Look Back and Predict Forward" improves image captions, providing significant insights. This research helps CV and NLP researchers and practitioners improve image captions by incorporating temporal factors.

Luo et al. [25] introduced the "Dual-Level Collaborative Transformer" to improve image captioning in this study. They used this method to create more contextually relevant and informative image captions.

In [26], the authors improved image captioning by using the caption better. They used caption information to create more contextually relevant and informative image captions.

In this study, the authors introduced "Attention on Attention" to improve image captioning [27]. They used this method to create more contextually relevant and informative image captions. This work helped computer vision researchers and practitioners understand how attention mechanisms might improve image captioning.

Liu et al. [28] presented a "Context-Aware Visual Policy Network" to improve sequence-level image

captioning. This method considered context and visual policies to provide more contextually relevant and informative image captions.

The authors introduced a "Recurrent Fusion Network" to improve image captioning in this study. Visual and textual information were repeatedly fused to create more contextually relevant and informative image captions [29]. It helped computer vision researchers and practitioners understand how recurrent fusion networks may improve image captioning.

Yao et al. [30] proposed using features to improve image captioning in this study. The goal was to use visual content attributes to create more contextually relevant and informative image captions. Hierarchy parsing was applied to improve image captioning in this study. The goal was to create more contextually appropriate and informative image captions by considering visual content hierarchy [31].

This study proposed distilling an image-text matching model to improve grounded image captions. Matching model findings were used to create more contextually relevant and informative image captions [32]. This study introduced a recall method for image captioning. A technique that emphasizes recalling pertinent information was used to create more contextually relevant and informative image captions [33].

A transformer network's intra- & inter-layer global representations were used to improve image descriptions in this study. This method considered global representations at several levels to create more contextually appropriate and informative image captions [34].

3. Evaluation Metrics

Evaluation metrics analyses the performance and quality of processes, systems, models, and algorithms. Image captioning and natural language processing use numerous evaluation measures to evaluate caption quality. These criteria assist researchers and practitioners evaluate image captioning methods. These image captioning metrics are common:

1. **BLEU (Bilingual Evaluation Understudy):** popular metric for assessing the calibre of text produced by machines, such as image captions, is called BLEU. Based on n-grams (word sequences), it calculates how similar the generated and reference captions are to each other. Greater quality captions are indicated by a higher BLEU score [35].
2. **ROUGE (Recall-Oriented Understudy for Gisting Evaluation):** The ROUGE set of measures is employed to assess the text quality produced by summarization systems. The measure concentrates on word recall and overlap between generated and reference captions [36].
3. **METEOR (Metric for Evaluation of Translation with Explicit ORdering):** Another measure that takes into consideration word overlap is METEOR, which also takes stemming, synonyms, and other linguistic variances into account. It provides a more thorough assessment of captions [37].
4. **CIDEr (Consensus-based Image Description Evaluation):** CIDEr was created especially for the assessment of image captions. It evaluates caption quality based on consensus from human annotators as well as word diversity and individuality [38].
5. **SPICE (Semantic Propositional Image Caption Evaluation):** One metric used to assess the semantic quality of image captions is called SPICE. It evaluates how well captions convey the visual content's meaning, taking into account object relationships [39].
6. **Perplexity:** The degree to which a language model accurately anticipates a given caption given the context supplied by the image is measured by its perplexity. Better model performance is indicated by lower confusion [40].
7. **METE (Metric for Evaluation of Textual Entailment):** METE evaluates how logically a image and its caption relate to one another. It assesses whether the caption supports or refutes the substance of the image [41].
8. **Human Evaluation:** Apart from automatic measurements, human assessment entails having hu-man annotators judge the calibre of created captions. This could involve standards like overall quality, relevancy, and fluency [42].

These assessment criteria are critical for evaluating the effectiveness of various image captioning models, fine-tuning model parameters, and monitoring industry advancements. For a comprehensive assessment of their image captioning systems, researchers frequently combine these measures.

4. Benchmark Datasets

To assess and compare the effectiveness of image captioning methods, benchmarking datasets are essential. They offer a set of uniform photos along with human-generated captions, enabling efficient comparisons between various models. An effective dataset can boost algorithm efficiency, Table 1 summarizes published datasets. These are a few benchmark datasets for image captioning:

1. **MS COCO (Common Objects in Context):** One of the most used benchmark datasets for image captioning is MS COCO. It has a vast assortment of unique photos, many of which have several human-annotated captions. The MS COCO benchmark is used to assess how well image captioning models perform. This is mostly made up of intricate scene images from Flickr, Yahoo's book-style website. Each image has five annotations, and there are 82,783 train images, 40 504 validation images, and 40,775 test images total [43].

2. **Flickr30K:** Another benchmark dataset, Flickr30K, has a sizable number of images from Flickr with five human-generated captions for each image. Research and assessment on image captioning frequently use it. The Flickr8k dataset has been expanded into Flickr30k. It has 31,783 images total, with five manual sentence labels assigned to each image [44].
3. **Flickr8K:** Like Flickr30K, a sizable portion of the images in the Flickr8K dataset have captions created by humans. It is commonly utilised as a benchmark dataset for jobs involving image captioning. The 8092 images in Flickr8k are divided into three categories: training (6092), verification (1000), and testing (1000). Five distinct sentences, each with an AVG length of 11.8 words, are tagged on each image. The dataset is manageable and appropriate for novice users [45].
4. **Visual Genome:** A substantial resource for vision-language challenges is the Visual Genome dataset. It is useful for assessing intricate image captioning models since it provides a large No.'s of images with thorough object and relationship annotations [46].
5. **AI2D (A Large-scale Dataset for Denotational Image Description):** The AI2D dataset offers a large-scale image and caption collection with an emphasis on denotational image description. Its purpose is to facilitate further research into this particular area of image captioning [47].

These benchmark datasets are used as a basis for comparing and developing image captioning models and are well-known in the field. These datasets are frequently used by researchers to test, train, and assess the effectiveness of their image captioning systems.

5. Experimental Evaluation

The Table 2 that is supplied provides an insightful comparison of several image captioning techniques according to their architectural elements and performance indicators. This data provides insightful information that image captioning researchers and practitioners can use to choose the best components and techniques for their particular applications. Improving the quality of spontaneously generated image descriptions and pushing the boundaries of image captioning require this kind of comparative investigation.

Table 1. Summarizes the performance of commonly accepted approaches. We give their accuracy score using conventional assessment measures on the MS COCO Karpathy-split-test set. They give demonstrations of their usage of attention mechanisms, visual encoding, and language decoding, and training strategies. Methods are grouped by suggested dates in Table 2.

In recent times, there has been a notable advancement in image captioning models. In terms of standard metrics, the B-4 score varies from 24.6 for global CNN features (NIC [2]) to 38.2 and 38.4 based on cross-entropy loss for graph encoding (CGVRG [26]) and self-attention encoding (X-LAN [18]), with a similar upward trend in reinforcement learning training.

The CIDEr score, which peaks at 140.4 for vision-and-language pre-training employing reinforcement learning, is absent from early grid feature models. It varies from 114.0 for region features to 135.6 for self-attention. We can further conclude that better captions are produced by structured and fine-grained visual semantic information as well as a variety of mutual interactions.

As opposed to Up-Down (fine-grained visual area features), GCN-LSTM [6] (organized visual information and relationships), and ETA [21], NIC [2] (coarse grid features) performs badly.

Reward learning can be a good substitute for cross-entropy loss, according to the results of several training methods. The most recent pre-training model, VinVL [9], finally has the highest ranking across all criteria.

Table 2. An overview of how many images were used in each dataset for testing, validation, and training. Additionally shown are the topic of the dataset and the number of caption labels for apiece image in the collection.

Dataset	Size			Captions/images	Topic
	Training	Validation	Testing		
Flickr8k [44]	6000	1000	1000	5	Human activities
Flickr30k [45]	28 000	1000	1000	5	Human activities
MSCOCO [43]	82 783	40 504	40 775	5	Daily scene
(Karpathy's split)	112 783	5000	5000	5	Daily scene

PASCAL 1K [48]	–	–	1000	5	Human activities
YFCC100M [49]	9920 million (32%)			7	Public multimedia
Multi30K-CLID [50]	29 000	1000	1000	5	Daily scene
AIC [51]	210 000	30 000	30 000 + 30 000	5	Daily scene
IAPR TC-12 [52]	17 665	–	196 2	1.7	Still natual
GoodNews [53]	424 000	18 000	23 000	1	News
VizWiz [54]	23 431	7750	8000	5	Blind view
Nocaps [55]	1 700 000	4500	10 600	10	Novel objects
FACAD [56]	993 000 images in total			0.2	Fashion items
TextCaps [57]	424 000	18 000	23 000	1	Text

Table 2. An overview of image captioning frameworks based on DL and their performances in various training approaches (attention Based). The Standard Metrics of Bleu-4, Meteor, Rouge-1, Cider, and Spice are represented by the markers "B4", "M," "R-L," "C," and "S," respectively. All results are derived from the Karpathy's-split of the MSCOCO dataset, and these scores are derived from the corresponding papers.

Method	**Encoder**	**Decoder**	**Cross-Entropy**					**Reinforcement Learning**				
			B-4	M	R-L	C	S	B-4	M	R-L	C	S
Soft-ATT [1]	CNN	LSTM	24.3	23.9	–	–	–	–	–	–	–	–
Hard-ATT [1]	CNN	LSTM	25.0	23.0	–	–	–	–	–	–	–	–
NIC [2]	CNN	LSTM	24.6	–	–	–	–	27.7	23.7	–	85.5	–
Adp-ATT [3]	CNN	LSTM	33.2	25.7	55.0	101.3	–	–	–	–	–	–
SCST [4]	CNN	LSTM	30.0	26.0	54.3	101.3	–	34.2	26.7	55.7	114.0	–
CPTR [5]	SA	T-ATT	–	–	–	–	–	40.0	29.1	59.4	129.4	–
GCN-LSTM [6]	GCN	LSTM	36.8	27.9	57.0	116.3	20.9	38.2	28.5	58.3	127.6	22.0
VSUA [7]	GCN	LSTM	–	–	–	–	–	38.4	28.5	58.4	128.6	22.0
MT [8]	SA	T-ATT	37.4	28.7	57.4	119.6	–	40.7	29.5	59.7	134.1	–
VinVL. [9]	SA	T-ATT	38.2	30.3	–	129.3	23.6	40.9	30.9	–	140.4	25.1
RSTNet [10]	SA	T-ATT	–	–	–	–	–	40.1	29.8	59.5	135.6	23.3
SGAE [11]	GCN	LSTM	–	–	–	–	–	38.4	28.4	58.6	127.8	22.1
CNM [12]	GCN	LSTM	37.1	27.9	57.3	116.6	20.8	38.7	28.4	58.7	127.4	21.8
Stack-Cap [13]	CNN	LSTM	35.2	26.5	–	109.1	–	36.1	27.4	56.9	120.4	20.9
GLA [14]	CNN	LSTM	31.2	24.9	53.3	96.4	–	–	–	–	–	–
Semantic-ATT [15]	CNN	LSTM	37.7	27.9	58.2	123.7	–	–	–	–	–	
DA [16]	CNN	LSTM	33.7	26.4	54.6	104.9	19.4	37.5	28.5	58.2	125.6	22.3
VRATT-Soft [17]	CNN	LSTM	34.3	28.5	60.0	111.7	20.1	37.5	28.5	61.6	122.1	22.1
VRATT-Hard [17]	CNN	LSTM	36.3	27.9	60.6	113.0	20.4	36.6	28.4	60.9	119.8	21.5
X-LAN [18]	SA	LSTM	38.2	28.8	58.0	122.0	21.9	39.5	29.5	59.2	132.0	23.4
X-T [18]	SA	T-ATT	37.0	28.7	57.5	120.0	21.8	39.7	29.5	59.1	132.8	23.4
ORT [19]	SA	T-ATT	35.5	28.0	56.6	115.4	21.2	38.6	28.7	58.4	128.3	22.6
NG-SAN [20]	SA	T-ATT	–	–	–	–	–	39.9	29.3	59.2	132.1	23.3
ETA [21]	SA	T-ATT	37.1	28.2	57.1	117.9	21.4	39.3	28.8	58.9	126.6	22.7
M2-T [22]	CNN	LSTM	–	–	–	–	–	39.1	29.2	58.6	131.2	22.6

OSCAR [23]	SA	T-ATT	36.5	30.3	−	123.7	23.1	40.5	29.7	−	137.6	22.8
LBPF [24]	CNN	LSTM	37.4	28.1	57.5	116.4	21.2	38.3	28.5	58.4	127.6	22.0
DLCT [25]	SA	T-ATT	−	−	−	−	−	39.8	29.5	59.1	133.8	23.0
CGVRG [26]	GCN	LSTM	38.4	28.2	58.0	119.0	21.1	38.9	28.8	58.7	129.6	22.3
AOANet [27]	SA	LSTM	36.9	28.5	57.3	118.5	21.6	39.1	29.0	58.9	128.9	22.5
CAVP [28]	CNN	LSTM	−	−	−	−	−	38.6	28.3	58.5	126.3	21.6
RFNet [29]	CNN	LSTM	35.8	27.4	56.8	112.5	20.5	36.5	27.7	57.3	121.9	21.2
LSTM-A [30]	CNN	LSTM	35.2	26.9	55.8	108.8	20.0	35.5	27.3	56.8	118.3	20.8
GCN-HIP [31]	GCN	LSTM	38.0	28.6	57.8	120.3	21.4	39.1	28.9	59.2	130.6	22.3
POS-SCAN [32]	CNN	LSTM	36.5	27.9	−	114.9	20.8	38.0	28.5	−	125.9	22.2
SRT [33]	SA	T-ATT	36.6	28.0	56.9	116.9	21.3	38.5	28.7	58.4	129.1	22.4
MAC [34]	SA	T-ATT	−	−	−	−	−	39.5	29.3	58.9	131.6	22.8

6. Conclusion

Within the fields of CV and NLP, automatic image captioning is a crucial field that focuses on converting visual input into text descriptions. Recent years have seen amazing progress in this interdisciplinary field, driven by the incorporation of DL techniques and the increasing availability of large-scale image-caption datasets. In this study, we perform a thorough overview of the history of image captioning, cover-ing a wide range of topics such as evaluation criteria, datasets, contemporary DL models, and traditional methods. We start by providing an overview of template-based and retrieval-based techniques, as well as their improvements. We then explore DL Image captioning models, emphasizing training methods, attention processes, and the encoder-decoder framework. We classify and condense assessment metrics and datasets that are pertinent to captioning images. We compare the performance of current approaches with the MS COCO benchmark and industry-standard evaluation metrics. Even while DL models have come a long way, they can still be improved. We wrap off by talking about some possible future routes for image captioning research. Among other fields, intelligent information transfer, smart homes, and education all depend heavily on image captioning. It is unquestionably important in the fields of DL and artificial intelligence, and in the future, its impact on our daily lives is predicted to increase. This analysis highlights the persisting hurdles and the opportunity for additional refining even though DL models for image captioning have made substantial progress. Discussion topics for future research are covered, with a focus on the significance of resolving problems with multi-modal representations, contextual comprehension, and fine-grained details.

References

1. K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in Proc. Int. Conf. Machine Learning, 2015, pp. 2048−2057
2. O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: Lessons learned from the 2015 MS COCO image captioning challenge,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol.39, no.4, pp.652–663, 2016.
3. J. Lu, C. Xiong, D. Parikh, and R. Socher, “Knowing when to look: Adaptive attention via a visual sentinel for image captioning,” in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition, 2017, pp. 375−383.
4. S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, “Selfcritical sequence training for image captioning,” in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition, 2017, pp. 7008−7024.
5. W. Liu, S. Chen, L. Guo, X. Zhu, and J. Liu, “CPTR: Full transformer network for image captioning,” arXiv preprint arXiv: 2101.10804, 2021.
6. T. Yao, Y. Pan, Y. Li, and T. Mei, “Exploring visual relationship for image captioning,” in Proc. European Conf. Computer Vision, 2018, pp. 684−699.
7. L. Guo, J. Liu, J. Tang, J. Li, W. Luo, and H. Lu, “Aligning linguistic words and visual semantic units for image captioning,” in Proc. 27th ACM Int. Conf. Multimedia, 2019, pp. 765−773.
8. J. Yu, J. Li, Z. Yu, and Q. Huang, “Multimodal transformer with multi-view visual representation for image captioning,” IEEE Trans. Circuits and Systems for Video Technology, vol.30, no. 12, pp. 4467– 4480, 2020.
9. P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao, “VINVL: Revisiting visual representations in vision-language models,” in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition, 2021, pp. 5579−5588.
10. X. Zhang, X. Sun, Y. Luo, J. Ji, Y. Zhou, Y. Wu, F. Huang, and R. Ji, “Rstnet: Captioning with adaptive attention on visual and non-visual words,” in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition, 2021, pp. 15465−15474.

11. X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-encoding scene graphs for image captioning," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition, 2019, pp. 10685−10694.
12. X. Yang, H. Zhang, and J. Cai, "Learning to collocate neural modules for image captioning," in Proc. IEEE/CVF Int. Conf. Computer Vision, 2019, pp. 4250−4260
13. J. Gu, J. Cai, G. Wang, and T. Chen, "Stack-captioning: Coarse-to-fine learning for image captioning," in Proc. AAAI Conf. Artificial Intelligence, 2018, vol. 32, no. 1, pp. 6837–6844
14. L. Li, S. Tang, L. Deng, Y. Zhang, and Q. Tian, "Image caption with global-local attention," in Proc. AAAI Conf. Artificial Intelligence, 2017, vol. 31, no. 1, pp. 4133−4239.
15. M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Paying more attention to saliency: Image captioning with saliency and context attention," ACM Trans. Multimedia Computing, Communications, and Applications, vol.14, no.2, pp.1–21, 2018.
16. L. Gao, K. Fan, J. Song, X. Liu, X. Xu, and H. T. Shen, "Deliberate attention networks for image captioning," in Proc. AAAI Conf. Artificial Intelligence, 2019, vol. 33, no. 1, pp. 8320−8327.
17. Z. Zhang, Q. Wu, Y. Wang, and F. Chen, "Exploring region relationships implicitly: Image captioning with visual relationship attention," Image and Vision Computing, vol. 109, p. 104146, 2021. DOI: 10.1016/j.imavis.2021.104146.
18. Y. Pan, T. Yao, Y. Li, and T. Mei, "X-linear attention networks for image captioning," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition, 2020, pp. 10971−10980.
19. S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning: Transforming objects into words," in Proc. Advances in Neural Information Processing Systems, 2019, pp. 11137−11147.
20. L. Guo, J. Liu, X. Zhu, P. Yao, S. Lu, and H. Lu, "Normalized and geometry-aware self-attention network for image captioning," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition, 2020, pp. 10327−10336.
21. G. Li, L. Zhu, P. Liu, and Y. Yang, "Entangled transformer for image captioning," in Proc. IEEE Int. Conf. Computer Vision, 2019, pp. 8928−8937.
22. M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed memory transformer for image captioning," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition, 2020, pp. 10578−10587.
23. X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, et al., "Oscar: Object-semantics aligned pre-training for vision-language tasks," in Proc. European Conf. Computer Vision, Springer, 2020, pp. 121−137
24. Y. Qin, J. Du, Y. Zhang, and H. Lu, "Look back and predict forward in image captioning," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition, 2019, pp. 8367−8375.
25. Y. Luo, J. Ji, X. Sun, L. Cao, Y. Wu, F. Huang, C.-W. Lin, and R. Ji, "Dual-level collaborative transformer for image captioning," in Proc. AAAI Conf. Artificial Intelligence, 2021, vol. 35, no. 3, pp. 2286−2293.
26. Z. Shi, X. Zhou, X. Qiu, and X. Zhu, "Improving image captioning with better use of caption," in Proc. 58th Annual Meeting Association for Computational Linguistics, 2020, pp. 7454−7464.
27. L. Huang, W. Wang, J. Chen, and X. Wei, "Attention on attention for image captioning," in Proc. IEEE Int. Conf. Computer Vision, 2019, pp. 4634−4643.
28. D. Liu, Z.-J. Zha, H. Zhang, Y. Zhang, and F. Wu, "Context-aware visual policy network for sequence-level image captioning," in Proc. 26th ACM Int. Conf. Multimedia, 2018, pp. 1416−1424.
29. W. Jiang, L. Ma, Y.-G. Jiang, W. Liu, and T. Zhang, "Recurrent fusion network for image captioning," in Proc. European Conf. Computer Vision, Springer, 2018, pp. 499−515.
30. T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," in Proc. IEEE/CVF Int. Conf. Computer Vision, 2017, pp. 4894−4902.
31. T. Yao, Y. Pan, Y. Li, and T. Mei, "Hierarchy parsing for image captioning," in Proc. IEEE/CVF Int. Conf. Computer Vision, 2019, pp. 2621−2629.
32. Y. Zhou, M. Wang, D. Liu, Z. Hu, and H. Zhang, "More grounded image captioning by distilling image-text matching model," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition, 2020, pp. 4777−4786.
33. L. Wang, Z. Bai, Y. Zhang, and H. Lu, "Show, recall, and tell: Image captioning with recall mechanism." in Proc. AAAI Conf. Artificial Intelligence, 2020, pp. 12176−12183.
34. J. Ji, Y. Luo, X. Sun, F. Chen, G. Luo, Y. Wu, Y. Gao, and R. Ji, "Improving image captioning by leveraging intra-and inter-layer global representation in transformer network," in Proc. AAAI Conf. Artificial Intelligence, 2021, vol. 35, no. 2, pp. 1655−1663.
35. Papineni, K., Roukos, S., Ward, T., & Zhu, W. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.
36. Lin, C. Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*.
37. 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*.
38. Vedantam, R., Zitnick, C. L., & Parikh, D. (2015). CIDEr: Consensus-based Image Description Evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
39. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2016). SPICE: Semantic Propositional Image Caption Evaluation. In *ECCV*.
40. Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. In *Computational Linguistics*, 19(2), 263-311.

41. Dolan, W. B., Quirk, C., & Brockett, C. (2004). Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. In *Proceedings of the 20th International Conference on Computational Linguistics*.
42. Elliott, D., & Keller, F. (2013). Image Description Using Visual Dependency Representations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.
43. Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision*.
44. Young, P., Lai, A., Hodosh, M., & Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. In *Transactions of the Association for Computational Linguistics*.
45. 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*.
46. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., ... & Fei-Fei, L. (2017). Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision, 123*(1), 32-73.
47. Sariyildiz, I. S., Yuret, D., & Karakaya, K. (2019). AI2D: A Large-scale Dataset for Denotational Image Description. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
48. M. Everingham, A. Zisserman, C. K. Williams, *et al.*, "The 2005 pascal visual object classes challenge," in *Proc. Machine Learning Challenges Workshop*, Springer, 2005, pp. 117−176.
49. B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "YFCC100m: The new data in multimedia research," *Commun. ACM*, vol. 59, no. 2, pp. 64–73, 2016.
50. D. Elliott, S. Frank, K. Sima'an, and L. Specia, "Multi30k: Multil- ingual english-german image descriptions," in *Proc. 5th Workshop on Vision and Language*, 2016, pp. 70−74.
51. J. Wu, H. Zheng, B. Zhao, Y. Li, B. Yan, R. Liang, W. Wang, S. Zhou, G. Lin, Y. Fu, Y. Z. Wang, and Y. G. Wang, "AI challenger: A large- scale dataset for going deeper in image understanding," arXiv preprint arXiv: 1711.06475, 2017.
52. M. Grubinger, P. Clough, H. Müller, and T. Deselaers, "The iapr TC-12 benchmark: A new evaluation resource for visual information systems," in *Proc. Int. Workshop OntoImage*, 2006, vol. 2, pp.13–23.
53. F. Biten, L. Gomez, M. Rusinol, and D. Karatzas, "Good news, everyone! Context driven entity-aware captioning for news images," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2019, pp. 12466−12475.
54. D. Gurari, Y. Zhao, M. Zhang, and N. Bhattacharya, "Captioning images taken by people who are blind," in *Proc. European Conf. Computer Vision*, Springer, 2020, pp. 417−434.
55. H. Agrawal, K. Desai, Y. Wang, X. Chen, R. Jain, M. Johnson, D. Batra, D. Parikh, S. Lee, and P. Anderson, "NOCAPS: Novel object captioning at scale," in *Proc. IEEE/CVF Int. Conf. Computer Vision*, 2019, pp. 8948−8957.
56. X. Yang, H. Zhang, D. Jin, Y. Liu, C.-H. Wu, J. Tan, D. Xie, J. Wang, and X. Wang, "Fashion captioning: Towards generating accurate descriptions with semantic rewards," in *Proc. Computer Vision-ECCV*, Springer, 2020, pp. 1−17.
57. O. Sidorov, R. Hu, M. Rohrbach, and A. Singh, "Textcaps: A dataset for image captioning with reading comprehension," in Proc. European Conf. Computer Vision, Springer, 2020, pp. 742−758