International Journal of Computer Engineering in Research Trends
ISSN: 2349-7084/ https://doi.org/10.22362/ijcert.v10i12.913
Volume 10, Issue 12, December 2023
© 2023, IJCERT All Rights Reserved

Research Paper

An Extensive Analysis of Image Captioning Models, Evaluation Measures, and Datasets

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: 18/10/2023, Revised: 21 /11/2023, Accepted: 19/12/2023 Published: 30/12/2023

Abstract: - *The difficult interdisciplinary endeavour of creating insightful and detailed captions for images is known as "image captioning," and it lies at the nexus of computer vision and natural language processing. We give a comprehensive examination of datasets, evaluation metrics, and image captioning models in this paper. We present a thorough review of popular image captioning models, from conventional methods to the most recent developments utilizing deep learning and attention mechanisms. We examine the design, underlying assumptions, and capabilities of these models, emphasizing how they help produce logical and contextually appropriate captions. Furthermore, we analyses in detail well-known evaluation metrics as BLEU, METEOR, ROUGE, and CIDEr, clarifying their importance in evaluating generated caption quality against ground truth references. Additionally, we talk about the critical role that datasets play in image captioning research, with particular attention to prominent datasets like as COCO, Flickr30k, and Conceptual Captions. We investigate these dataset's diversity, volume, and annotations, emphasizing their impact on model evaluation and training. Our objective is to furnish scholars, professionals, and amateurs with an invaluable tool for comprehending the state of image captioning, so facilitating the creation of inventive models and enhanced assessment techniques.*

Keywords- *Image Caption, language template, Search-Based, supervised learning, Visual Space, Deep learning.*

1. Introduction

In the fields of computer vision and natural language processing, image captioning—the automatic generation of meaningful and descriptive textual descriptions for images—has drawn a lot of interest. A variety of image captioning models have been developed as a result of the convergence of language generation and visual understanding. These models use advanced approaches to close the semantic gap between verbal and visual information. Assessing the effectiveness and caliber of these models is essential, which is why specific assessment metrics that gauge the degree of similarity between generated captions and ground truth annotations are employed. Moreover, the performance and efficacy of these models are shaped during training and benchmarking by the availability of a wide range of well-annotated datasets. This study undertakes a thorough investigation of image captioning models, covering their designs, innovations, and distinctive characteristics. It explores the assessment metrics and offers an understanding of how these measurements affect the evaluation of generated captions. This paper also examines a number of image captioning datasets, emphasizing their importance and influence on the development of image captioning research. We hope that this thorough analysis will give scholars and practitioners a thorough grasp of image captioning techniques, allowing them to navigate this dynamic field and spur new breakthroughs.

Emphasizing the semantic relationship between visual information and corresponding natural language terms, the Farhadi et al. [1] provided an automated method for producing textual descriptions for photographs. By teaching models to recognize this association, their method makes it possible to produce meaningful and cogent words that explain the information contained in the images.

Advanced models that bridge the gap between visual perception and language interpretation were made possible by this groundbreaking study, which served as the basis for later image captioning research.

Combines emotive themes drawn from both linguistic and visual elements in a novel way for captioning images. Yang et al. [2] suggested a technique that integrates sentiment- and emotion-based indicators into photos in order to capture affective information. The model attempts to produce captions that are not only descriptive but also infused with emotive awareness, so augmenting the emotional context of the created words, by merging visual elements and textual data. This work showcases the potential to enhance image descriptions with emotional nuances, and it marks a substantial exploration into the merging of affective computing with image captioning.

In paper [3] authors presented an in-depth analysis of many image captioning models, examining their designs, processes, and distinctive characteristics. The study also emphasizes the value of various datasets and how they are used to train and assess image captioning models. It also explores several assessment metrics, emphasizing how to apply and evaluate them to judge the calibre and coherence of output captions. As a valuable resource, this assessment summarizes the most recent methods for image captioning and offers suggestions for future developments in this ever-evolving subject.

The Bernardi et al. [3] given a thorough summary of all the many methods, approaches, and strategies used in image captioning. With an emphasis on the transition from conventional to contemporary deep learning-based techniques, the survey discusses the development of image captioning models. By highlighting their advantages and disadvantages, it examines the parts and architectures of these models. The study also explores the assessment metrics used to gauge the effectiveness of captioning models and includes a discussion of datasets essential to image captioning research. For the purpose of comprehending the state of automatic image caption generation and the direction of future developments, this survey is a useful resource for instructors and researchers.

Important contributions of "An Extensive Analysis of Image Captioning Models, Evaluation Measures, and Datasets" include:

Summarizing numerous research papers and studies on image captioning and offering a thorough assessment of the state of the subject at present time.

Thoroughly evaluating and comparing various evaluation metrics commonly used to measure the quality of generated captions, shedding light on their effectiveness.

Offering practical advice and guidelines for practitioners, aiding them in choosing suitable models, evaluation metrics, and datasets for their image captioning projects.

Reviewing and assessing popular datasets used in image captioning, discussing their relevance and impact on model training and evaluation.

2. Literature Review

The language template-based approach and the search-based method are the two previous approaches that are introduced in this section.

2.1 Language Template-Based Approach

The language template-based approach to image captioning is a structured and rule-driven method for generating descriptive captions for images. In this approach, predefined sentence templates or structures are created, designed to accommodate relevant image-related information. These templates act as a scaffold for constructing captions and guide the process to ensure consistency and coherence. The templates often include placeholders or slots where specific image details, such as objects, scenes, or actions identified within the image, can be inserted.

The process involves extracting information from the image, typically through computer vision techniques like Convolutional Neural Networks (CNNs), and placing the extracted details into the appropriate slots in the templates. This systematic insertion of image-related information into the templates results in the generation of captions that provide a clear and structured description of the image content. While the language template-based approach offers a structured and controlled way of generating captions, it may face challenges in handling the rich variability and nuances present in natural language compared to more advanced methods like those utilizing deep learning models.

Yang et al. [4] suggest an approach that uses a corpus of sentences as a guide to help create descriptive sentences that match real-world visuals. With the help of this corpus, the model is able to learn and infer linguistic structures and patterns, which helps to produce meaningful and contextually relevant descriptions for images. By using linguistic insights and patterns from a large corpus, this research advances the field of image captioning and produces more accurate, logical, and human-language-like captions.

In order to shed light on words or phrases that are unexpected in a particular context, Dunning [5] investigates methods for calculating measures of surprise for them. The study presents the widely used statistical measure known as "Log-Likelihood Ratio (LLR)," a technique used in a number of natural language processing tasks, including sentiment analysis, information retrieval, and language modelling. With his foundational paradigm for statistical analysis in computational linguistics and related fields, Dunning's contributions have had a major impact on the discipline.

2.2 Search-Based Image Captioning Approach

Search-based image captioning retrieves suitable textual descriptions or captions from a dataset or collection that matches the image. This method uses similarity metrics or information retrieval to find the most relevant captions from a dataset. The selection is based on the closeness of image features extracted from the given image and dataset caption features. Captions are used to describe images. This method uses existing annotated datasets to solve problems quickly and efficiently, making it beneficial when real-time generation or prolonged training is not possible. It depends largely on the dataset's quality and diversity and may miss the image's originality or details.

In [6] create a model that generates image textual descriptions using the rich source of annotated images and captions. By using the vast volume of image-caption pairs, this work advanced image captioning and showed the promise of data-driven image interpretation and natural language production. The "Im2Text" method has shaped image captioning research and methods.

A method to rank prepared written descriptions based on visual relevance is proposed by the Hodosh et al.[7]. Novel ranking-based evaluation criteria and datasets are introduced. By making image description a ranking problem, this study improves evaluation and illuminates how to improve image caption quality and relevance, advancing image understanding and natural language generation.

2.3 Deep Learning Methods

Image captioning has been revolutionized by deep learning. Models combine trans-formers or RNNs for captioning and CNNs for image characteristics. Attention mechanisms concentrate on important aspects of the image. Many captions are made possible using variational autoencoders. Repurposed are pretrained models such as BERT and GPT. GANs improve variation and realism. These techniques transform the process of creating insightful visual descriptions.In the context of sequence learning, Lipton, Berkowitz, and Elkan [8] offered a thorough evaluation of recurrent neural networks (RNNs). It highlights the uses and difficulties faced by RNNs in a range of domains while critically analyzing their bene-fits and drawbacks. The review provides insightful information about the capabilities and future directions of RNNs in sequence-related tasks.

Zaremba, Sutskever, and Vinyals [9] concentrate on methods for recurrent neural networks (RNNs) regularization. It presents techniques to improve RNN training and generalization, offering important insights into regularization strategies for this particular kind of neural network.

The study investigated the incorporation of compositional semantics to link descriptive phrases with visuals. It explores how images are found and described using linguistic compositions, providing a thorough grasp of how language may be utilized to interact with visual content. This study opens the door for more advanced techniques in image understanding and description by taking a key step forward in the interaction between language and images [10].

Karpathy, Joulin, and Fei-Fei [11] introduced an innovative approach using deep fragment embeddings for bidirectional mapping between images and sentences. By representing images and sentences as embeddings, the study explores methods to establish meaningful connections and associations between visual and linguistic data. This work contributes to advancing the understanding of how deep learning techniques can facilitate bidirectional understanding between images and textual descriptions, demonstrating the potential for improved image-sentence mapping.

Lebret, Pinheiro, and Collobert [12], presented at the International Conference on Machine Learning, proposes a phrase-based approach to generate captions for images. This method employs a structured prediction framework to generate phrases in a compositional manner, forming descriptive captions for images. By introducing a novel way of constructing captions using phrases, the study contributes to enhancing the quality and interpretability of image descriptions, offering a significant advancement in image captioning techniques.

Kiros, Salakhutdinov, and Zemel [13] introduced and talked about multimodal neural language models. The objective of these models is to enhance language generation and comprehension by combining data from several modalities, such text and visuals. Through combining textual and visual data, the research investigates new methods for improved language modelling and representation learning. This study makes a substantial contribution to the multimodal learning field in general as well as to our understanding of the intricate interactions that exist between various data types.

The study [14] investigates the significance and impact of incorporating explicit high-level concepts in solving vision-to-language problems. By exploring the role of these concepts in various tasks at the intersection of vision and language, the re-search sheds light on their contribution and relevance in improving performance and understanding in this domain. This work is pivotal in understanding the importance of high-level concepts in bridging the gap between visual data and natural language.
Karpathy and Fei-Fei [15] focused on leveraging deep learning techniques to align visual and semantic representations for generating image descriptions. By integrating visual and semantic information, the study aims to improve the quality and relevance of automatically generated image captions. The research contributes to advancements in image understanding and description by demonstrating the effective-ness of aligning deep visual and semantic features.

2.4 Region Based Approach

Liu et al. [16] announced the Swin Transformer, a hierarchical vision transformer using shifted windows. By hierarchically splitting the image into non-overlapping patches or windows and permitting communication between them, this strategy im-proves vision transformer efficiency and efficacy. Vision transformer models can handle huge image inputs better with the Swin Transformer's structured and efficient processing technique.

Guidance mechanisms improve Long-Short Term Memory (LSTM) image caption creation in the study. The project attempts to increase caption quality and relevancy by using image data signals. This study is crucial to improving LSTM models' contex-tual image descriptions [17]. In the paper [18] presented DenseCap, which generates dense image captions for localization and captioning. The method describes image regions using fully convolutional networks. This work advances visual interpretation by allowing fine-grained captioning.

2.5 Novel object

Yao et al. [19] used a hierarchical system to analyses photos, making captions more organized. The research uses this hierarchical strategy to improve image caption coherence and meaning. This study advances new image caption creation methods. The study [20] added a pointing feature that lets the model point to novel things in the image and generate captions. The research improves image captioning by addressing novel objects, especially when standard models struggle to describe them. This study illuminates image captioning with different and unseen things.

In [21] authors used guiding artefacts to make captions more relevant and intentional. The research focuses on intentions and guiding objects to improve visual content-caption alignment and create more meaningful and intentional image descriptions. This research provides insights into writing captions that convey the image's message.

2.6 Visual Space vs. Multimodal Space

In image captioning, "Visual Space" centers on image features extracted using techniques like CNNs. On the other hand, "Multimodal Space" integrates both image features and linguistic context, aiming to generate more contextually relevant and descriptive captions by aligning visual and linguistic information.

2.7 Supervised Learning Image Captioning

Supervised learning in image captioning involves training a model using labeled image-caption pairs [22]. Images are processed through a CNN to extract features, which are then used by an RNN, like LSTM, to generate captions. The model learns to generate accurate captions by minimizing the difference between predicted and actual captions using cross-entropy loss.

2.8 Encoder-Decoder Architecture-Based Image Captioning

This encoder, usually based on CNNs, examines the image and extracts high-level characteristics that encapsulate its visual content. The decoder, commonly based on RNNs or transformers, derives the caption from these attributes. A meaningful and contextually relevant caption is generated by training the decoder to predict the next word in the sequence using encoder features. Encoder Decoder design has successfully bridged image visual information and natural language descriptions. The encoder learns to extract key image attributes that the decoder uses to create a cohesive caption [23].

3. Benchmark Datasets

The most often used large-scale dataset in image captioning is ***MS COCO [24],*** which mostly includes complex scene images from Flickr. It contains 82,783 train images, 40,504 validation images, and 40,775 test images with 5 annotations each.

Flickr8k [25] has 8,092 images, with 6,092 for training, 1,000 for verification, and 1,000 for testing. The images are labelled with 5 statements averaging 11.8 words each. Small datasets are good for novices.

Flickr30k [26] expands flickr8k. The database has 31,783 images with 5 hand sentence labels apiece. This dataset has 94.2% human images, 12% animal images, 69.9% clothing images, 28% limb movement images, and 18.1% automobiles and other instruments. It lacks fixed image segmentation for training, testing, and validation.

PASCAL 1K [27] expands beyond object detection to image captioning. The database features 20 categories, each with 50 randomly selected photographs and 5 private captions. All photographs are from Flickr.

YFCC100M [28] has 99.2 million Yahoo Flickr images, with 32% having captions, average 7 sentences per image and 22.52 words each sentence.

Multi30K-CLID [29] expands Flickr30K [26] to include a multilingual description dataset with 29,000 training photos, 1,014 verification images, and 1,000 test images. Description languages: English, German, French, and Czech. Each language offers 5 image annotations.

AIC [30] is the first Chinese language caption dataset. Every image in the collection was gathered through online searches. Over 200 sceneries and 150 actions are included, with 210,000 photos for training, 30,000 for verification, and 60,000 for testing. Five Chinese annotations are included for each image.

IAPR TC-12 [31] includes over 20,000 images from various sites worldwide, with most donated by venture. Images often have an average of 1.7 words of annotations. The dataset has 17,665 training photos.

GoodNews [32] is the largest news caption collection comprised of articles from 2010-2018. The database has 466,000 photos with handwritten captions, headlines, and text articles. The data is divided into three sets: 424,000 for training, 18,000 for validation, and 23,000 for random testing. In particular, GoodNews annotations are produced by professional journalists rather than crowd-sourced.

Nocaps [33] is the first large-scale benchmark for novel object captioning. The benchmark includes 166,100 hand-marked captions for 15,100 images from Open Images V4

[37] validation and test sets. The validation and testing split has three subsets: in-domain, near-domain, and out-of-domain, corresponding to different levels of COCO nearness.

The innovative fashion dataset **FACAD [34]** aims to examine captioning for fashion items. The collection of over 993K fashion photographs from Google Chrome spans various seasons, ages, genres, and body orientations. About 130K descriptions, averaging 21 words, are pre-processed for future research.

The TextCaps **[35]** collection provides 28k photos from Open images v3 with 145k captions to understand text in images. The dataset tests the model's ability to recognize text, contextualize it visually, and choose which parts to reproduce or interpret. It performs spatial, semantic, and visual reasoning between text tokens and visual elements.

Visual Genome [36] contains 108,077 images from MS COCO [24] and YFCC [156], with 5.4 million region descriptions, 3.8 million object instances, 2.8 million characteristics, and 2.3 million associations. The typical image has 35 items, 26 attributes, and 21 pairwise interactions between objects. It is commonly used for model pre-training in relational learning-based image description studies.

VizWiz [38] is for blind image captioning services. The database has 31981 photos of blind individuals, each with 5 captions. The split is approximately 70%/10%/20% for train/val/test.

4. Evaluation Metrics

BLEU (Bilingual Evaluation Understudy) [39] is a key statistic for assessing machine-generated text in natural language processing applications like image captioning. BLEU's main function is to compare generated text to human-reference material, typically several translations. The algorithm analyses n-grams, contiguous sequences of n elements, usually words. BLEU measures the amount of overlapping n-grams in the generated text compared to the reference to determine the adjusted precision for each n-gram and calculates the overall BLEU score using a geometric mean.

Higher BLEU scores indicate greater similarity between created and reference texts. The downsides of BLEU include not addressing semantic equivalence, fluency, or coherence and being sensitive to slight text changes. BLEU is a frequently used statistic for evaluating generated text in natural language processing applications, including image captioning, despite its drawbacks.

METEOR (Metric for Evaluation of Translation with Explicit ORdering) [40] stands as a critical metric in the realm of machine-generated text evaluation, prominently used in assessing the quality of translations and generated text across a spectrum of natural language processing applications, including image captioning. Unlike BLEU, METEOR incorporates linguistic and semantic features in its assessment. It operates by comparing the generated text against one or more human-authored reference texts. This comparison involves measuring overlaps using unigrams, stemming to handle morphological variations, and considering synonyms through the use of resources like WordNet.

The METEOR score, which ranges from 0 to 1, is calculated by combining precision and recall, giving more weight to recall. The metric aims for a balanced assessment by encouraging precision while penalizing excessive word generation. METEOR's ability to account for both lexical and semantic similarities makes it a valuable tool in evaluating the quality and accuracy of generated text in various natural language processing tasks, contributing significantly to the advancement of the field.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [41] is a set of metrics commonly used for evaluating the quality of machine-generated text, particularly in the field of natural language processing. The primary purpose of ROUGE is to assess the similarity and overlap between the generated text and one or more reference texts. It measures this similarity by considering n-grams, word sequences of varying lengths. ROUGE calculates several metrics, including ROUGE-N (which considers unigrams, bigrams, trigrams, and more), ROUGE-L (which focuses on the longest common subsequence), ROUGE-W (which considers word overlap), and others. These metrics evaluate the precision and recall of the n-grams, providing insights into how well the generated text aligns with the reference. The ROUGE score is calculated by combining precision and recall, with higher scores indicating a better match between the generated and reference texts. ROUGE is extensively utilized in machine translation, summarization, and image captioning, offering a quantitative measure of the quality and relevance of the generated text, and aiding researchers and practitioners in refining and optimizing natural language processing models.

CIDEr (Consensus-based Image Description Evaluation) [42] is a significant evaluation metric in the domain of image captioning, devised to measure the quality and relevance of the generated captions. Unlike metrics like BLEU or METEOR, CIDEr places a stronger emphasis on consensus and consensus-based evaluation. It leverages a consensus mechanism that accounts for human agreement by pooling multiple reference captions for each image. CIDEr evaluates the generated caption by comparing it against these references, calculating a consensus score that signifies the agreement in word usage between the generated and reference captions.

CIDEr operates by analyzing n-grams, which are contiguous sequences of n words, typically up to n=4. It computes a term frequency-inverse document frequency (TF-IDF) based similarity, wherein TF-IDF values for each n-gram are calculated and then aggregated to produce the final CIDEr score. Higher CIDEr scores indicate a greater consensus and alignment with human references, implying a more accurate and relevant generated caption.

This metric plays a crucial role in evaluating the performance of image captioning models, as it provides a quantitative measure of how well the model-generated captions agree with human-generated ones. CIDEr is widely adopted due to its ability to consider the diversity and consensus in human annotations, offering a more comprehensive and insightful assessment of the quality of image captions generated by automated systems.

SPICE (Semantic Propositional Image Caption Evaluation) [43] is a popular image captioning metric that evaluates semantic content and propositions. SPICE prioritizes text interpretation and semantic representation over language structures. It compares generated captions against human-authored reference captions to see if they represent the intended meaning and the image's object-action linkages. SPICE deconstructs captions into semantic objects, properties, and relationships. It uses a parser to arrange captions into semantic graphs for more extensive analysis.

The SPICE score measures semantic content agreement by assessing these semantic networks' similarity and overlap. By considering meaning and semantics, SPICE improves on existing measures like BLEU and METEOR by providing a more nuanced view of caption quality. It is commonly used to evaluate image captioning systems and develop and modify models that provide more accurate and contextually meaningful captions.

5. Flicker & COCO Image Captioning

In computer vision and natural language processing, the task of image captioning is to produce a descriptive sentence or phrase that appropriately captures the content of the image. In this area, one of the most popular datasets is COCO, or Common Objects in Context. It may be used to train and assess image captioning models because it has images with thorough annotations. Whereas "Flicker" is referring to the Flickr dataset, then this is another dataset that is frequently utilized for comparable objectives and offers a wide range of images (See Table 1).

Researchers usually use a variety of deep learning models in the image captioning field using the MS COCO dataset (See Table 2). These models are often based on convolutional neural networks (CNNs) for image feature extraction and recurrent neural networks (RNNs) or transformers for caption generation. The common procedures and things to keep in mind when comparing methods are summarized as follows:

Yao et al. [50] and Fan et al. [51] achieved the highest BLEU-1 score, with 0.809 and 0.818, respectively. Fan et al. [51] outperformed others in BLEU-4 with a score of 0.408. Karpathy et al. [15] and Jia et al. [44] demonstrated competitive performance across all BLEU scores. Fan et al. [51] had the highest METEOR score of 0.295, indicating good overall performance. Yao et al. [50] also showed a notable METEOR score of 0.286.

Fan et al. [51] had the highest CIDEr score of 1.353, suggesting better agreement with human consensus. Yao et al. [50] also demonstrated a high CIDEr score of 1.287. Yao et al. [50] had a SPICE score of 0.221, indicating its ability to capture semantic content in captions. Fan et al. [51] had a SPICE score of 0.225, showing similar competence.

Chen et al. [58] stood out in terms of METEOR and CIDEr scores, showcasing its alignment with human judgments. Fan et al. [51] consistently demonstrated strong performance across various metrics, making it a notable candidate for effective image captioning. Dong et al. [53] and Yao et al. [57] also performed consistently well across multiple metrics.

The attention mechanism and Transformer-based approaches appear to be at the forefront of this study area. Deep learning-based methods are the primary focus of current research in this sector. The attention mechanism is taught to filter image regions and direct sentence fragments for image captioning in the Lu et al. [49] technique.

The best performance on Flcikr8k is shown by the BLEU-1, BLEU-2, BLEU-3, and BLEU-4 ratings, which are, respectively, 0.677, 0.494, 0.354, and 0.251. Additionally, ROUGE scores—0.531—are supplied for this dataset.

Pan et al.'s [52] attention mechanism-based image captioning technique obtained superior scores on the Microsoft COCO Caption dataset. Current captioning encoders use graph convolutional networks (GCN) to represent object relationships in the query image, which may explain the exceptional performance of the attention mechanism. Image objects and scenes are represented by deep CNNs.

The attention mechanism is effective for CNN, GCN, and sequence-based decoders, creating an ideal framework. Adding attention to the encoder-decoder framework can improve caption performance by conditioning phrase production on hidden states estimated using the attention technique.

Attention should also be paid to the rising direction of novelty entities. We found that this technique performs better on smaller Flickr datasets. Fu et al. [48] and Lu et al. [49] improved performance on the Flcikr30k dataset. The scores for BLEU-1, BLEU-2, BLEU-3, and BLEU-4 are 0.720, 0.459, 0.319, and 0.285, respectively.

The CIDEr score of 0.648 surpasses other methods. It was found that captions created by existing methods only capture training goals and cannot be applied to novel settings or objects, leading to poor performance. Thus, selecting new objectives and expanding descriptive phrase vocabulary is interesting. However, the Transformer-based technique has achieved the highest scores on BLEU-1, BLEU-2, and ROUGE (0.822, 0.676, and 0.597 on MS COCO dataset) as a rising star.

The RNN's inability to handle long-distance dependency may cause gradients to evaporate or explode after numerous propagation stages.

Deep learning offers advantages over RNN in engineering due to its representation learning and ability to expand to various layers. These factors restrict the advancement of this framework. Powerful visual encoders and attention-guided image captioning can compete with Transformer-based approaches in performance. While slower to train, these approaches are typically smaller than Transformer-based ones.

The Transformer-based technique addresses RNN long-distance dependence. Its structure makes it easy to add layers for layout purposes. We anticipate additional researchers creating innovative Transformer-based designs.

6. Future Research Directions Image Caption

It is expected that further image captioning research will propel developments in a number of important areas. Multimodal techniques will become more popular, combining data from different modalities—such as text, audio, and images—to produce captions that are richer and more thorough. Another emphasis is on precise captioning of images, with the goal of capturing minute visual details. Context awareness will be essential in generating captions that are more cohesive and pertinent by taking into account the larger context.

The ability of dynamic attention mechanisms to adaptively focus on particular items or regions of an image, imitating human attention patterns, will be improved. In order to enable models to produce meaningful captions for previously unseen categories or with few instances, zero-shot and few-shot learning will be investigated. Captions that are more in line with logical, human-like descriptions will result from the integration of visual commonsense reasoning into image captioning models.

Video captioning will be expanded to include both spatial and temporal contexts through the use of spatiotemporal understanding. To lessen reliance on substantial annotated datasets and enable more effective training, self-supervised learning techniques will be studied. There will be a greater emphasis on ethical issues, such as eliminating prejudices and guaranteeing equity in all demographic circumstances. The ability for users to customize generated captions will improve the user experience through user-centric customization. Moreover, modifying image captions for languages with limited resources will promote diversity and assist a wider spectrum of linguistic communities.

All things considered, these lines of inquiry are critical to expanding the possibilities of image captioning and improving its inclusivity, contextual awareness, and practicality.

Table 1. Method comparison on datasets Flcikr8k and Flickr30k

Method	Datasets	B-1	B-2	B-3	B-4	M	R	C
Karpathy et al. [15]	Flcikr30k	0.573	0.369	0.240	0.157	-	-	-
	Flcikr8k	0.579	0.383	0.245	0.160	-	-	-
Jia et al. [44]	Flcikr30k	0.647	0.459	0.318	0.216	-	0.202	-
	Flcikr8k	0.646	0.446	0.305	0.206	-	0.179	-
Lee et al. [45]	Flcikr30k	0.274	-	-	0.286	0.403	0.300	0.419
	Flcikr8k	-	-	-	-	-	-	-
Xu et al. [46]	Flcikr30k	0.670	0.457	0.314	0.213	0.203	-	-
	Flcikr8k	0.669	0.;439	0.296	0.199	0.185	-	-
Lu et al. [47]	Flcikr30k	-	-	-	-	-	-	-
	Flcikr8k	0.677	0.494	0.354	0.251	0.204	0.531	-
Fu et al. [48]	Flcikr30k	0.639	0.459	0.319	0.217	0.204	0.470	0..538
	Flcikr8k	0.649	0.462	0.324	0.224	0.194	0.451	0.472
Lu et al. [49]	Flcikr30k	0.720	-	-	0.285	0.231	-	0.648
	Flcikr8k	-	-	-	-	-	-	-

Table 2. Method comparison on MS COCO dataset under the commonly used protocol

Method	B-1	B-2	B-3	B-4	M	R	C	S
Karpathy et al. [15]	0.625	0.450	0.321	0.230	—-	0.195	0.660	—-
Jia et al. [44]	0.670	0.491	0.358	0.264	0.227	—-	0.813	—-
Yao et al. [50]	0.809	—-	—-	0.383	0.286	0.585	1.287	0.221
Fan et al. [51]	0.818	—-	—-	0.40.8	0.295	0.592	1.353	0.225
Xu et al. [46]	0.718	0.504	0.357	0.250	0.230	—-	—-	—-
Lu et al. [47]	0.742	0.580	0.439	0.332	0.266	—-	1.085	—-
Pan et al. [52]	0.817	0.668	0.526	0.407	0.299	0.597	1.353	0.238
Dong et al. [53]	0.822	0.676	0.524	0.398	0.298	0.597	1.294	—-

Ji et al. [54]	0.816	0.665	0.519	0.397	0.294	0.591	1.303	—-
Liu et al. [55]	0.818	0.665	0.518	0.395	0.291	0.592	1.254	—-
Lu et al. [49]	0.759	—-	—-	0.349	0.274	—-	1.089	0.201
Yang et al. [56]	0.810	0.656	0.507	0.385	0.282	0.586	1.238	0.222
Yao et al. [57]	0.816	0.662	0.515	0.393	0.288	0.590	1.279	0.223
Chen et al. [58]	—-	—-	—-	0.230	0.245	0.501	2.042	0.421
Mahajan et al. [59]	0.777	0.620	0.473	0.354	0.270	0.563	1.144	0.204
Liu et al. [60]	0.754	0.591	0.445	0.332	0.257	0.550	1.013	—-
Wu et al. [61]	0.790	0.630	0.482	0.363	0.277	0.571	1.179	0.216
Deng et al. [62]	0.776	0.613	0.469	0.353	0.286	0.574	1.182	0.223
Mao et al. [63]	0.670	0.490	0.350	0.250	—-	—-	—-	—-
Vinyals et al. [64]	0.666	0 .461	0.329	0 .246	—-	—-	—-	—-
Wang et al.[64]	0.672	0.492	0.352	0.244	—-	—-	—-	—-
Wang et al. [65]	0.687	0.509	0. 364	0.258	0.229	—-	0.739	—-
Donahue et al. [66]	0.697	0.519	0.380	0.278	0.229	0 .508	0.837	—-
Xu et al. [46] (Soft att)	0.707	0.492	0.344	0.243	0.239	—-	—-	—-
You et al. [67]	0.709	0 .537	0.402	0.304	0.239	—-	—-	—-
Aneja et al. [68]	0.693	0 .518	0.374	0.268	0.238	0. 511	0.855	—-
Li et al.[69]	0.718	0.543	0.395	0.286	0.242	0.52 3	0.91 2	—-
Huei et al. [70]	0.706	0.534	0. 395	0.292	0. 23 7	0.517	0.881	—-
Zhu et al. [71]	0.733	0 .570	0 .436	0.333	-	0.548	1.081	—-

7. Conclusion

The comprehensive analysis of image captioning models, evaluation measures, and datasets provides valuable insights into the rapidly evolving field of computer vision and natural language processing. The study encompassed an in-depth exploration of various models, ranging from traditional methods to cutting-edge deep learning architectures, elucidating their strengths, weaknesses, and specific applications. The critical evaluation of evaluation measures shed light on their efficacy in assessing the quality and relevance of generated captions, offering a nuanced under-standing of their suitability for diverse contexts. Additionally, the thorough review of datasets, considering their size, diversity, and annotations, unveiled their impact on model performance and highlighted the necessity of appropriate dataset selection for robust image captioning.

This extensive analysis not only serves as a benchmark for understanding the cur-rent landscape of image captioning but also lays the foundation for future research directions. The identified gaps and potential areas for improvement pave the way for advancements that could revolutionize image captioning, fostering innovations in multimodal integration, fine-grained details, context awareness, and ethical considerations. As image captioning continues to play a vital role in applications like accessibility, content summarization, and human-computer interaction, this analysis sets the stage for the development of more effective and inclusive image captioning systems, ultimately benefiting a wide array of users and domains.

References

1. Farhadi, A., Hejrati, M., Sadeghi, M.A., et al.: Every picture tells a story: Generating sentences from images. In: European Conference on Computer Vision, pp. 15–29. (2010)
2. Yang, J., Sun, Y., Liang, J., et al.: Image captioning by incorporating affective concepts learned from both visual and textual components. Neurocomputing (2019), 328, 56–68
3. Bernardi, R., Cakici, R., Elliott, D., et al.: Automatic description generation from images: A survey of models, datasets, and evaluation measures. J. Artif. Intell. Research (2016), 55, 409–442
4. Yang, Y., et al.: Corpus-guided sentence generation of natural images. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. (2011)
5. Dunning, T.E.: Accurate methods for the statistics of surprise and coincidence. Comput. Linguist. 19(1), 61–74 (1993)
6. Ordonez, V., Kulkarni G., Berg, T.: Im2text: Describing images using 1 million captioned photographs. Adv. Neural Inf. Process. Syst. 24, 1143– 1151 (2011)
7. Hodosh, M., Young P., Hockenmaier, J.: Framing image description as a ranking task: Data, models and evaluation metrics. J. Artif. Intell. Research 47, 853–899 (2013)
8. Lipton, Z.C., Berkowitz, J., Elkan, C.: A critical review of recurrent neural networks for sequence learning. Comput. Sci. abs/1506.00019, (2015).

9. Zaremba, W., Sutskever I., Vinyals, O.: Recurrent neural network regularization. arXiv:1409.2329 (2014)
10. Socher, R., et al.: Grounded compositional semantics for finding and describing images with sentences. Trans. Assoc. Comput. Linguist. 2, 207–218 (2014)
11. Karpathy, A., Joulin, A., Fei-Fei, L.: Deep fragment embeddings for bidirectional image sentence mapping. In: Proceedings of the Twenty Seventh Advances in Neural Information Processing Systems (NIPS), vol. 3, pp. 1889–1897 (2014)
12. Lebret, R., Pinheiro P., Collobert R.: Phrase-based image captioning. In: International Conference on Machine Learning (2015)
13. Kiros, R., Salakhutdinov, R., Zemel, R.: Multimodal neural language models. In: International Conference on Machine Learning, pp. 595–603 (2014)
14. Wu, Q., Shen, C. & Liu, L.: What value do explicit high level concepts have in vision to language problems? In: IEEE conference on computer vision and pattern recognition, pp. 203-212 (2016)
15. Karpathy, A. & Fei-Fei L. : Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference On Computer Vision And Pattern Recognition (2015)
16. Liu, Z., Lin, Y., Cao, Y., et al.: Swin transformer: Hierarchical vision transformer using shifted windows. arXiv:2103.14030 (2021)
17. Jia, X., Gavves E., Fernando B. & Tuytelaars T. : Guiding the long-short term memory model for image caption generation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2407–2415 (2015)
18. Johnson, J., Karpathy A., Fei-Fei L.,: Densecap: Fully convolutional localization networks for dense captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4565–4574 (2016)
19. Yao, T., et al.: Hierarchy parsing for image captioning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2019)
20. Li, Y., et al.: Pointing novel objects in image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)
21. Zheng, Y., Li Y., Wang S.: Intention oriented image captions with guiding objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8395–8404 (2019)
22. Johnson, J., Karpathy, A., & Fei-Fei, L. (2016). DenseCap: Fully Convolutional Localization Networks for Dense Captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4565-4574.
23. Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and Tell: A Neural Image Caption Generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3156-3164. DOI: 10.1109/CVPR.2015.7298935
24. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in Proceedings of the European conference on computer vision. Springer, 2014, pp. 740–755.
25. P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, “From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions,” Transactions of the Association for Computational Linguistics, vol. 2, pp. 67–78, 2014.
26. B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, “Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models,” in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2641–2649.
27. M. Everingham, A. Zisserman, C. K. Williams, L. Van Gool, M. Allan, C. M. Bishop, O. Chapelle, N. Dalal, T. Deselaers, G. Dorkó et al., “The 2005 pascal visual object classes challenge,” in Machine Learning Challenges Workshop. Springer, 2005, pp. 117–176.
28. 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,” Communications of the ACM, vol. 59, no. 2, pp. 64–73, 2016.
29. D. Elliott, S. Frank, K. Sima’an, and L. Specia, “Multi30k: Multilingual english-german image descriptions,” in Proceedings of the 5th Workshop on Vision and Language, 2016, pp. 70–74.
30. J. Wu, H. Zheng, B. Zhao, Y. Li, B. Yan, R. Liang, W. Wang, S. Zhou, G. Lin, Y. Fu et al., “Ai challenger: A large-scale dataset for going deeper in image understanding,” arXiv preprint arXiv:1711.06475, 2017.
31. M. Grubinger, P. Clough, H. Müller, and T. Deselaers, “The iapr tc-12 benchmark: A new evaluation resource for visual information systems,” in International workshop ontoImage, vol. 2, 2006.
32. A. F. Biten, L. Gomez, M. Rusinol, and D. Karatzas, “Good news, everyone! context driven entity-aware captioning for news images,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 12 466–12 475.
33. 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 Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 8948–8957.
34. 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 Computer Vision–ECCV 2020. Springer, 2020, pp. 1–17.
35. O. Sidorov, R. Hu, M. Rohrbach, and A. Singh, “Text caps: a dataset for image captioning with reading comprehension,” in Proceedings of the European Conference on Computer Vision. Springer, 2020, pp. 742–758.
36. R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma et al., “Visual genome: Connecting language and vision using crowdsourced dense image

annotations," International journal of computer vision, vol. 123, no. 1, pp. 32–73, 2017.

37. I. Krasin, T. Duerig, N. Alldrin, V. Ferrari, S. Abu-ElHaija, A. Kuznetsova, H. Rom, J. Uijlings, S. Popov, A. Veit et al., "Open images: A public dataset for largescale multi-label and multi-class image classification," Dataset available from https://github. com/open images, vol. 2, no. 3, p. 18, 2017.
38. D. Gurari, Y. Zhao, M. Zhang, and N. Bhattacharya, "Captioning images taken by people who are blind," in Proceedings of the European Conference on Computer Vision. Springer, 2020, pp. 417–434.
39. K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.
40. S. Banerjee and A. Lavie, "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, vol. 29, 2005, pp. 65–72.
41. C. Lin and F. J. Och, "Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics," in Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, 2004, pp. 605–612.
42. R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2015, pp. 4566– 4575.
43. P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in Proceedings of the European Conference on Computer Vision. Springer, 2016, pp. 382–398.
44. Jia, X., Gavves E., Fernando B. & Tuytelaars T. : Guiding the long-shortterm memory model for image caption generation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2407–2415(2015)
45. Lee, H., Yoon, S., Dernoncourt, F., et al.: UMIC: An unreferenced metric for image captioning via contrastive learning. arXiv:2106.14019 (2021)
46. Xu, K., et al.: Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning(2015)
47. Lu, J., Xiong, C., Parikh, D., et al.: Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)(2017)
48. Fu, K., Jin, J., Cui, R., et al.: Aligning where to see and what to tell: Image captioning with region-based attention and scene-specific contexts. IEEE Trans. Pattern Anal. Mach. Intell. 39(12), 2321–2334(2016)
49. Lu, J., Yang, J., Batra, D., et al.: Neural baby talk. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 7219–7228 (2018)
50. Yao, T., et al.: Exploring visual relationship for image captioning. In: Proceedings of the European Conference On Computer Vision (ECCV)(2018)
51. Fan, Z., Wei, Z., Wang, S., et al.: TCIC: Theme concepts learning cross language and vision for image captioning. arXiv:2106.10936 (2021)
52. Pan, Y., Yao T., Li Y., Mei T.: X-linear attention networks for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10971–10980 (2020)
53. Dong, X., Long, C., Xu, W., et al.: Dual graph convolutional net-works with transformer and curriculum learning for image captioning.arXiv:2108.02366 (2021)
54. Ji, J., Luo, Y., Sun, X., et al.: Improving image captioning by leveraging intra-and inter-layer global representation in transformer network. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35,pp. 1655–1663 (2021)
55. Liu, W., Chen, S., Guo, L., et al.: CPTR: Full transformer network for image captioning. arXiv:2101.10804 (2021)
56. Yang, X., et al.: Auto-encoding scene graphs for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)
57. Yao, T., et al.: Hierarchy parsing for image captioning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2019)
58. Chen, S., et al.: Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
59. Mahajan, S., Roth, S.: Diverse image captioning with context-object split latent spaces. arXiv:2011.00966 (2020)
60. Liu, S., Zhu Z., Ye N., Guadarrama S., Murphy K.: Improved image captioning via policy gradient optimization of spider. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 873–881 (2017)
61. Wu, J., Chen, T., Wu, H., et al.: Fine-grained image captioning with global-local discriminative objective. IEEE Trans. Multimedia 23, 2413–2427(2021)
62. Deng, C., Ding, N., Tan, M., et al.: Length-controllable image captioning. In: Computer Vision–ECCV 2020: 16th European Conference, pp. 712–729. Glasgow, UK, 23–28 August 2020
63. Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. "Deep captioning with multimodal recurrent neural networks (m-rnn)". In: *International Conference on Learning Representations (ICLR)*. 2015.
64. Cheng Wang, Haojin Yang, Christian Bartz, and Christoph Meinel. "Image captioning with deep bidirectional LSTMs". In: *Proceedings of the 2016 ACM on Multimedia Conference*. ACM. 2016, pp. 988–997.
65. Cheng Wang, Haojin Yang, and Christoph Meinel. "Image captioning with deep bidirectional lstms and multi-task learning". In: *ACM Transactions on*

Multimedia Computing, Communications, and Applications (TOMM) 14.2s (2018), p. 40.

66. Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. "Long-term recurrent convolutional networks for visual recognition and description". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 2625–2634.
67. Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. "Image captioning with semantic attention". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 4651–4659.
68. Jyoti Aneja, Aditya Deshpande, and Alexander G Schwing. "Convolutional image captioning". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018,pp. 5561–5570.
69. Linghui Li, Sheng Tang, Lixi Deng, Yongdong Zhang, and Qi Tian. "Image caption with global-local attention". In: *Thirty-First AAAI Conference on Artificial Intelligence* (2017).
70. Jia Huei Tan, Chee Seng Chan, and Joon Huang Chuah. "COMIC: Towards a compact image captioning model with attention". In: *IEEE Transactions on Multimedia* (2019).
71. Xinxin Zhu, Lixiang Li, Jing Liu, Haipeng Peng, and Xinxin Niu. "Captioning transformer with stacked attention modules". In: *Applied Sciences* 8.5 (2018), p. 739.