A Survey on Available Methods in Image Caption Generator

Dr. J V Gorabal[1], Anirudh Nitin Bakare[2], Daniel D[3], Kishore K[4], Manjunatha D[5]

[1]Professor, Department of CSE, ATMECE, Mysuru
[2,3,4,5]Student, Dept of CSE, ATMECE, Mysuru

---***---

Abstract - Image caption generation is a vital research area in the fields of computer vision and natural language processing. Over the years, various approaches have been utilized to generate captions for images by utilizing deep learning techniques, computer vision models, and NLP algorithms. However, evaluating the quality of the captions generated and integrating external knowledge into the process still pose a challenge. In this survey paper, we present a comprehensive review of the different approaches used for generating image captions and assess their performance. We also discuss the applications of image caption generation and the current challenges associated with the process. Moreover, we provide valuable insights into potential future research directions in this area. Our survey paper's objective is to furnish researchers and practitioners with a comprehensive understanding of cutting-edge image caption generation techniques and prospective research opportunities.

Keywords: CNN, RNN, LSTM, Image, Caption, Deep Learning

1. Introduction

The process of creating an accurate and comprehensive verbal depiction of the content contained within an image, which is commonly referred to as a caption, is referred to as image caption generation. The generation of image captions is a significant research area, which combines computer vision and natural language processing subfields. There have been several advancements in deep learning techniques, computer vision models, and natural language processing algorithms, which have resulted in various applications in different fields such as multimedia content understanding, image and video recognition, and human-computer interaction. However, incorporating external knowledge into the captioning process, as well as evaluating the quality of the generated captions, still poses a significant challenge to researchers, who continue to work on enhancing the accuracy, relevance, and diversity of the image captioning models.

Several alternate techniques have been proposed for addressing image description issues, such as the query expansion technique introduced by Yagcioglu et al. [1], which involves retrieving analogous images from an extensive dataset and utilizing the distribution explained in conjunction with the retrieved images. This expression is utilized to develop an expanded query, followed by reordering the candidate descriptions by approximating the cosine amid the distributed representation and the extended query vector. Lastly, the nearest description is a depiction of the input image. In summary, the approaches are inventive and possess distinct attributes, but all share a common disadvantage of not making intuitive feature observations on objects or actions in the image, nor do they offer an end-to-end comprehensive universal model to overcome this predicament. The effectiveness and popularity of neural networks have resulted in breakthroughs in the field of image description, offering new prospects with the onset of the big data era and emergence of deep learning techniques.

2. Extracting features of an image

In the realm of image caption models, there exist two primary categories: Firstly, a technique that employs a statistical probability language model for generating handcrafted features; and secondly, a neural network model founded on an encoder-decoder language model to extract deeper features.

2.1 Handcraft Features with Statistical Language Model

The Midge system utilizes a technique of maximum likelihood estimation to acquire the language model and visual detector directly from the image description dataset. This approach, depicted in Figure 1, was initially studied by Fang et al. [2], where in the picture is first examined and the object is identified to generate a caption. To identify words, a convolutional neural network (CNN) is employed on the image region [3], and the details are merged with MIL [4]. Then, the sentence structure is trained straight from the caption to decrease any prior assumptions about it. Ultimately, the system transforms the image caption generation into an optimization problem and recognizes the most probable sentence.

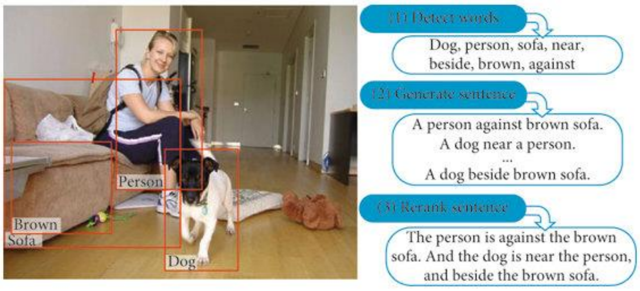
Person
Brown
Sofa
Dog
(1) Detect words
Dog, person, sofa, near, beside, brown, against
(2) Generate sentence
A person against brown sofa.
A dog near a person.
...
A dog beside brown sofa.
(3) Rerank sentence
The person is against the brown sofa. And the dog is near the person, and beside the brown sofa.

***Fig 1*: AI visual detector in conjunction with a language model**

The steps to be taken to implement is:

1. The algorithm identifies a group of words that could potentially be included in the image caption. The model employs a weak monitoring approach in multi-instance learning (MIL) to detect relevant vocabulary based on the content of the corresponding image, thereby facilitating iterative detector training.

2. When implementing a fully convolutional network on an image, a spatial response map is produced, whereby each position represents a response obtained from executing the CNN on various regions of the image via shifting. This procedure facilitates the ability to detect feasible objects present within the image. By up sampling the image, a response map is generated for the conclusive fully connected layer, prompting the use of the noisy-OR version of the Multiple Instance Learning (MIL) technique. This application to each image's response map culminates in the computation of probabilities for each corresponding word.

3. The caption generation process involves searching for the most probable sentence while considering the visually identified set of words. At the core of this process lies the language model, which defines the probability distribution for a sequence of words. Such information can facilitate the identification of incorrect words and the encoding of common-sense knowledge.

4. Numerous techniques may be utilized for image caption generation, including attribute detectors and language models. In their work, Devlin et al. [5] employed a combination of CNN and k-NN methods along with a maximum entropy model and RNN to generate descriptions for images. Kenneth Tran [6] proposed a similar system that utilizes a CNN as a visual model capable of detecting various visual concepts, landmarks, celebrities, and entities. The visual features are then integrated into a language model, and the resulting vectors are utilized as input to a multichannel depth-similar model for generating descriptions.

2.2 Deep Learning Features with Neural Network

The Recurrent Neural Network (RNN) [7] has attracted significant interest within the realm of deep learning due to its promising results in natural language processing, particularly in language modelling [8]. RNNs are also utilized for speech-to-text and text-to-speech conversion [9-10], machine translation, and question and answer sessions. As powerful language models at both character and word levels, RNNs have also gained popularity in computer vision. The widely used encoder-decoder model for image description generation employs RNN through an encoder, which is a Convolutional Neural Network (CNN) that extracts image features from the last layer's convolutional features, and a decoder, which generates image descriptions using RNN. However, training RNNs can be challenging [11], as it is subject to the general gradient descent problem. While regularization may mitigate this issue [12], RNNs have a limitation they can only remember the previous unit's content for a limited time, which can be addressed by utilizing LSTM. With long-term memory and a solution to the gradient vanishing problem, LSTM has shown promising results in handling video-related contexts in recent years [13].

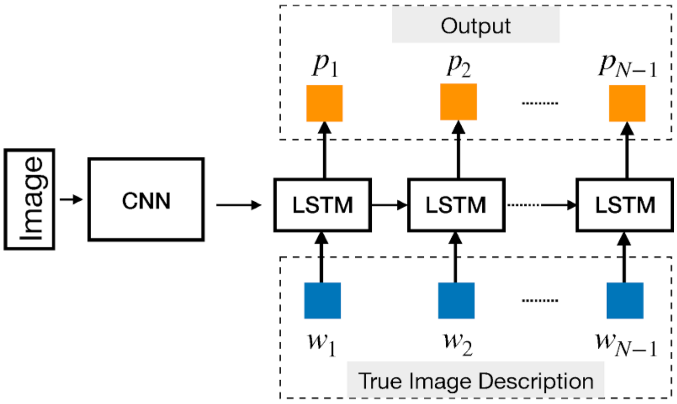

***Fig 2*:** Model based on encoder-decoder.

3. Attention Mechanism

The attention mechanism is a pivotal element of image caption generators. By utilizing this mechanism, the model is able to concentrate on the most relevant components of the image, resulting in an accurate and detailed caption. The mechanism operates by assigning a weight to each part of the image, based on its relevance to the captions content. Throughout the captioning process, the model repeatedly selects and combines various parts of the image, along with their corresponding weights, to construct a more comprehensive and accurate representation of the image that is subsequently employed to generate the final caption.

This methodology facilitates the production of captions that are not only detailed and informative, but also aligned with the human intuition regarding effective image descriptions. The attention mechanism constitutes a noteworthy advancement in image captioning, which has measurably enhanced the precision and quality of the resultant captions. Therefore, it has expanded the scope of applications where such automated image labelling and tools for visually impaired individuals can be employed.

The attention mechanism employed in neural network models enables the selective concentration on a subset of inputs or features whilst disregarding others. This mechanism ensures the specific targeting of inputs or features and comprises two integral components: firstly, identifying the section of the input necessitating attention, and secondly, allocating the limited information processing resources to the critical components. Categorically, attention mechanisms can be classified into the following groups:

Soft Attention [36] is typically utilized in machine translation, where it calculates the weighted sum of an entire region of an image. By computing a weighted attention vector, a deterministic model can be established. The parameterization of soft attention allows for its embedding and modelling, thereby enabling its direct training. Gradient can be passed back through the attention mechanism module to other parts of the model.

Hard Attention [36]. Rather than calculating a weighted sum of all regions like soft attention, hard attention concentrates on one specific spot by randomly choosing a distinctive location. This approach of probabilistically sampling the hidden state of the input, instead of the entire encoder's hidden state, mandates Monte Carlo sampling to approximate the module's gradient for backpropagation. However, one disadvantage of hard attention lies in the method of selecting information based on maximum or random sampling, resulting in an inability to achieve a functional relationship between the final loss function and the attention distribution, thus hindering the use of backpropagation algorithm for training.

Multi Head Attention. The fundamental concept of global attention [36] is to incorporate the hidden layer state of all encoders in the computation. To derive the attention weight distribution, the current hidden layer state of the decoder is compared with each hidden layer state of the encoder. During the decoding process, like soft attention, the attention weight for each word in the encoding must be calculated at each time step, and subsequently, weighted by the context vector.

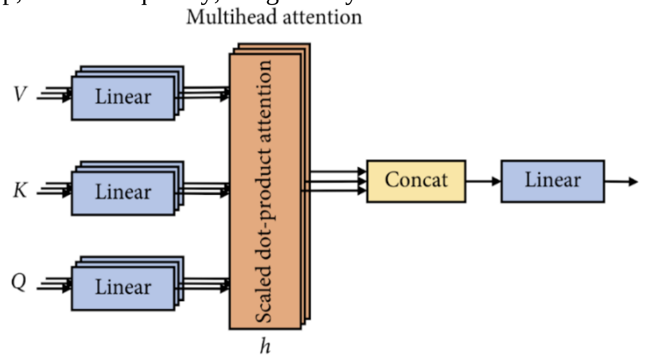

Fig 3: Multihead attention.

Global Attention. The primary concept of global attention [36] involves consideration of the hidden layer state of all encoders. By comparing the present decoder hidden layer state to each encoder hidden layer state, it obtains the attention weight distribution. This approach is like soft attention, where the attention weight of each word in the encoding must be calculated at each time step during decoding, followed by context vector weighting. The overall flow of this process is illustrated in Figure 4. However, as the computation of each decoder state requires consideration of all encoder inputs, the amount of calculation involved is relatively high.

Local attention [36]. initially identifies the alignment position and subsequently computes the attention weight in the left and right windows where the position exists, followed by context vector weighting. This methodology represents a compromise between soft and hard attention mechanisms. The primary benefit of local attention is that it reduces the cost involved in computing attention. During calculation, local attention need not consider all words on the source language side; instead, it predicts the alignment position of the source-language end for the current decoding through a prediction function and navigates through the context window, considering solely the words within the window.

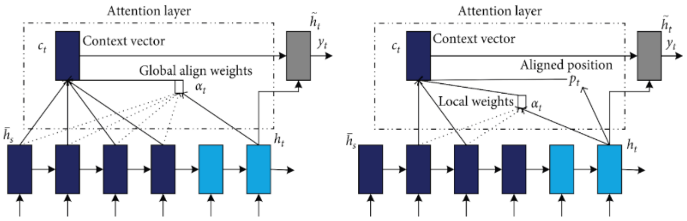

Fig 4: (a) Global attention model and (b) local attention model

Adaptive Attention. In image captioning and visual question answering, most attention models [36] adopt a strategy whereby the image is focused on at each time step, no matter which word is generated next. However, not all words correspond to visual signals. The adaptive attention mechanism and the visual sentinel address this problem by determining when and where to incorporate attention mechanisms to extract relevant information for each sequential word. As illustrated in Figure 5, the context vector represents the residual visual information of the LSTM hidden state, reducing uncertainty and supplementing information for the accurate prediction of the next word in the present hidden state.

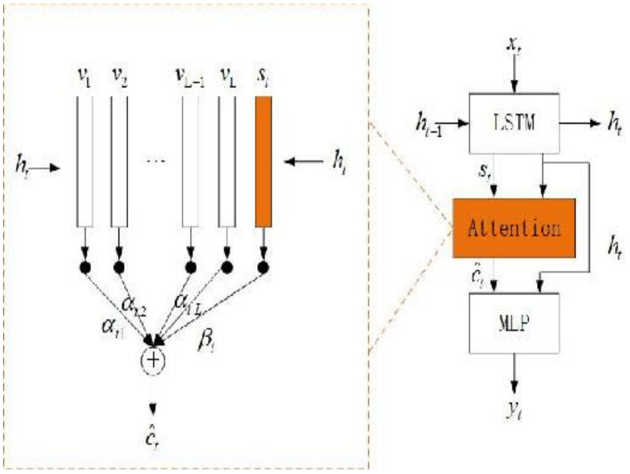

Fig 5: Adaptive attention

Spatial and Channel-Wise Attention. Spatial and channel attention is a process that selects semantic attributes based on the requirements of sentence context, as depicted in Figure 6. This technique utilizes attention mechanisms to overcome limitations associated with the decoding stage. The attention mechanism is applied on extracted semantics from the encoding process to prevent the overfitting problem typically associated with using the last layer. Attention is applied in multiple layers, as the feature map depends on its underlying feature extraction, thus obtaining visual attention in diverse semantic abstractions.

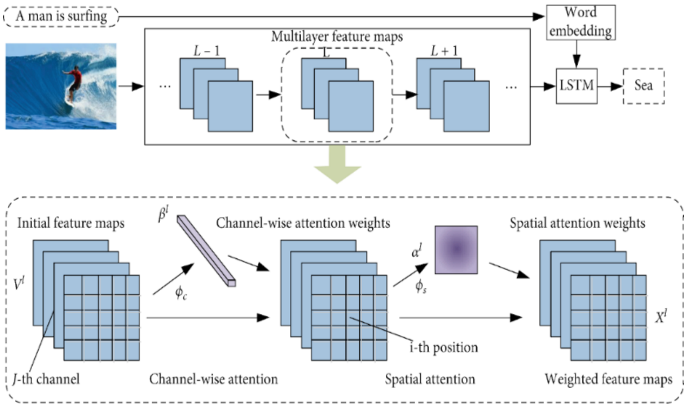

Fig 6: SCA-CNN model.

This chapter provides an analysis of various algorithm models for attention mechanisms. Table 1 presents a summary of attention mechanism applications in image description and outlines the feedback on different attention mechanisms and their modelling approaches. This feature enables readers to select an appropriate model for future research conveniently. The use of attention mechanisms results in an improved model effect.

Table 1: Comparison of attention mechanism modelling methods.

Attention name	Method	Comment
Soft attention	Gives a probability according to the context vector for any word in the input sentence when. seeking attention probability distribution	Parameterization Derivative enable Definitely
Hard attention	Focus only on a randomly chosen location using Monte Carlo sampling to estimate the gradient	Randomly On the base of probability Simple
Multihead attention	Linearly projecting multiple pieces of information selected from the input in parallel using multiple keys, values, and queries	Linear projection Parallel Focus on information from different representation subspaces in different locations Multiple attention head
Global attention	Considering the hidden layer state of all encoders, the weight distribution of attention is obtained by comparing the current decoder hidden layer state with the state of each encoder hidden layer	Comprehensive Time-consuming Large amount of calculation
Local attention	First find a location for it, then calculate the attention weight in the left and right windows of its location, and finally weight the context vector	Reduce the cost of calculations
Adaptive attention	Define a new adaptive context vector which is modelled as a mixture of the spatially attended image features and the visual sentinel vector. 0is trades off how much new information the network is considering from the image with what it already knows in the decoder memory	Solve when and where to add attention to extract meaningful information for sequence words
Spatial and channel-wise attention	Select semantic attributes based on the needs of the sentence context	Multiple semantics to overcome the problem of over range when using the general attention

4. Dataset and Evaluation

This section thoroughly examines commonly utilized datasets and metrics in current studies of image captioning, which have been implemented with varying approaches. However, due to the relatively restricted scope of available datasets for this task in contrast to object detection, and the inherent limitations of the evaluation metrics employed, further efforts must be made to expand the datasets and attain more precise evaluation metrics. It is vital to prioritize the development of more extensive and accurate datasets and metrics as the importance of image captioning in the industry continues to escalate, facilitating the progress and advancement of this task.

4.1 Datasets

There are several datasets accessible for creating image caption generators, including frequently used sets like MSCOCO, Flickr30k, Flickr 8k, AIC and Visual Genome. These sets comprise vast groups of images accompanied by corresponding captions and are extensively utilized in research in this domain. Nevertheless, it is crucial to thoughtfully analyze the distinct features and stipulations of each dataset before picking one for a specific project.

MSCOCO Captions dataset [14], which was created by the Microsoft Team and focuses on understanding scenes, captures images from intricate daily scenes, enabling multiple tasks such as image recognition, segmentation, and description.To produce descriptive captions for every image, the dataset utilizes Amazon's "Mechanical Turk" service, wherein a minimum of five sentences are generated for each image, resulting in a total of over 1.5 million captions. The dataset comprises of 82,783 images in the training set, 40,504 images in the validation set, and 40,775 images in the test set. Its 2014 version possesses roughly 20G images and around 500M in annotation files that correlate to an image and its descriptions.

Flickr30k dataset [15] is a collection of images used for captioning tasks. With approximately 30,000 images, it is one of the largest datasets available for this purpose. A unique characteristic of this dataset is that it includes five captions for each image, resulting in a total of about 150,000 captions. This enables the creation and training of captioning models with a wide range of approaches, such as text-based retrieval and generation-based methods. The Flickr30k dataset has become a standard benchmark for image captioning research due to its large size and its usefulness for evaluating the performance of captioning models.

Flickr 8k dataset [16]. is a collection of images intended for use in image captioning research. The dataset comprises approximately 8,000 images, each with five captions that describe different aspects of the image. The captions were written by human annotators and are concise, informative, and descriptive. The Flickr 8k dataset has become a popular benchmark for evaluating the accuracy and effectiveness of computer vision and natural language processing models that generate captions for images. Due to its relatively modest size compared to other similar datasets, it is often used as a starting point for new research in this field.

PASCAL 1K [17] is a segment of the renowned PASCAL VOC Challenge image dataset that offers a standardized image annotation dataset and assessment mechanism. The PASCAL VOC photo collection consists of 20 categories, and for its 20 categories, 50 images were randomly selected for a total of 1,000 images. Then, Amazon’s Turkish robot service is used to manually mark up five descriptions for each image. The dataset image quality is good, and the label is complete, which is very suitable for testing algorithm performance.

AIC. A Chinese Image Description dataset is a collection of images and matching textual descriptions in Chinese that is designed to support research in the field of image captioning. The dataset comprises over 210,000 images, each with five different human-written descriptions in Chinese. The images are diverse, covering a wide range of subjects, and were gathered from various sources, including social media platforms and online image databases. AIC Chinese Image Description dataset has allowed researchers to train and test machine learning models to generate accurate and informative image descriptions in Chinese. The dataset is a valuable resource for advancing research in natural language processing and computer vision.

Visual Genome. Unlike the other dataset discussed which only had one caption for the entire image, this dataset [18] presents a separate caption for each image region. This dataset comprises seven parts: region descriptions, attributes, relationships, region graphs, scene graphs, and question-answer pairs. The Visual Genome dataset contains more than 108 thousand images, with each image having an average of 35 objects, 26 attributes, and 21 pairwise relationships between objects.

Table 2: Summary of the number of images in each dataset

Dataset name	Train	Size Valid	Test
MSCOCO	82783	40504	40775
Flickr30k	28000	1000	1000
Flickr8k	6000	1000	1000
PASCAL 1K	—	—	1000
AIC	210000	30000	30000

4.2 Evaluation Metrics for Image Captioning Methods

The metrics presented can be categorized into two groups: text evaluation metrics and caption evaluation metrics. Text evaluation metrics are used to independently evaluate text generated by machines, primarily for assessing the quality of translations. Alternatively, caption evaluation metrics are employed to evaluate the quality of machine-generated captions and have been specifically developed for image captioning tasks.

BLEU (Bilingual Evaluation Understudy) [19]. evaluates machine-generated text by comparing parts of the text to reference texts and giving each part a score, then taking the average of these scores. BLEU counts consistent n-grams in both texts, with "n" determining the grams compared. BLEU has advantages like being language-independent, simple, fast, and comparable to human judgment, but disadvantages like unreliable high scores for short texts and sometimes indicating low-quality text.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [20]. rates text summaries by comparing them to ideal human summaries with overlapping units like n-grams, word sequences, & pairs. Metrics: ROUGE-N (overlapping n-grams), ROUGE-L (based on longest subsequence), ROUGE-W (weighted longest subsequence), ROUGE-S (skip-bigram co-occurrence), & ROUGE-SU (skip-bigram and unigram co-occurrence).

METEOR (Metric for Evaluation of Translation with Explicit Ordering) [21]. compares word segments to reference texts via harmonic mean of unigram precision and weighted recall. Has features such as stemming and synonymy matching. Better correlation at sentence/segment level.

CIDEr (Consensus-based Image Description Evaluation) [22]. Assessing the coherence of image annotations involves utilizing a TF-IDF weight computation for every n-gram present within sentences which are regarded as "documents". It utilizes cosine similarity to measure the likeness of machine-generated and reference descriptions, while also compensating for a drawback of BLEU by attributing greater importance to significant terms. A heightened score indicates superior performance.

Table3: Scores of attention mechanisms based on the evaluations above.

Ref	Attention model	BLEU-4	METEOR	ROUGE-L	CIDEr
[37]	Soft attention	24.3	23.9	—	—
[37]	Hard attention	25.0	23.0	51.6	86.5
[38]	Multihead	28.4	—	—	—
[39]	Global/local attention	25.9	—	—	—
[40]	Adaptive Attention	33.2	26.6	55.0	108.5
[41]	Spatial and channel-wise	31.1	25.4	53.0	94.3

The outcomes of the attention mechanisms outlined in section 3 are presented in Table 3. The data depicted in this table suggest that the effectiveness of diverse models differs depending on the evaluation criteria applied. However, if the attention model significantly improves performance, then its ranking is generally high across all evaluation metrics, despite some variations in specific evaluation measures.

Xu et al. [24] present techniques for producing captions utilizing attention mechanisms, enhancing the NIC model [23] as the existing leading method. The proposed techniques entail two types of attention mechanisms, namely a "hard" attention mechanism and a "soft" attention mechanism. The triumphs in caption generation lately and the profitable implementation of attention mechanisms in domains like machine translation [25] and object recognition [26] have prompted this approach, where the models intend to concentrate on the salient segments of an image while writing captions.

The primary methods for converting images to text are categorized as either top-down or bottom-up approaches. Top-down methods begin with a general idea of the image and subsequently generate corresponding words, while bottom-up approaches create descriptive phrases for various features of the image before merging them. To reconcile these two methods, You et al. [27] have developed a novel algorithm utilizing a semantic attention model. This innovation involves learning to focus on semantic concept proposals and combining them into the hidden states and outputs of recurrent neural networks by means of selective attention. The selection and combination process establishes a feedback loop that enhances the connection between top-down and bottom-up computations. This method has demonstrated marginally better performance than both "hard" and "soft" attention mechanisms.

Visual attention models traditionally emphasize spatial attention, but Chen et al. [28] have introduced a novel convolutional neural network known as SCA-CNN that combines spatial and channel-wise attentions within a single CNN. This advancement allows SCA-CNN to dynamically regulate the context for sentence generation across multilayer feature maps, capturing both the location and content of visual attention. Pedersoli and Lucas [27] propose an "Areas of Attention" technique which models the interdependencies between image regions, caption words, and the state of an RNN language model. By employing three pairwise interactions, this method establishes a direct association between caption words and image regions. The application of both these methods jointly has achieved previously unprecedented results on the MSCOCO dataset.

Lu et al. [29] have presented an adaptive attention model which features a visual sentinel. The model assesses whether to give precedence to the image or visual sentinel, and accurately directs attention to retrieve noteworthy data to be utilized for generating sequential words. This methodology has exhibited superior performance, surpassing the previous state-of-the-art and setting a new benchmark.

5. Challenges and the Future Directions

While multiple proposed solutions and methods have been suggested to address image captioning, some challenges and open problems persist. The effectiveness of supervised methods relies heavily on the quality of the datasets. However, even with extensive datasets, they may fail to encompass the entirety of the real world, and supervised methods are confined to the objects that the detector is trained to differentiate. The supervised paradigm overly relies on the language priors, which can lead to the object-hallucination phenomenon as well [30].

The limitations of supervised methods have prompted researchers to explore unsupervised techniques. However, due to the disparate properties of image and text modalities, the encoders of image and sentence cannot be shared. Thus, in an unpaired setting, the critical challenge is the information misalignment gap between images and sentences, resulting in lower performance rankings for current unsupervised image captioning methods [31]. Scene graphs have emerged as a promising direction for research into image captioning, revealing many possibilities as discussed in previous sections, but their utilization also presents challenges. Constructing scene graphs is a complex task, and the interactions between objects extend beyond simple pairwise relations, making integration of scene graphs a tedious process [32]. Additionally, scene graph parsers are still less potent [33]. Studies examining the impact of scene graphs on caption quality indicate that pre-training of the scene graph generators with visually relevant relation data is crucial to their effectiveness [34].

VLP methods have been employed to address some of the deficiencies associated with supervised methods and object detector-based designs. However, while VLP approaches are well-suited for comprehension tasks, the generation of tasks such as image captioning requires additional capabilities. This gap has been the focus of numerous recent studies reviewed in this paper; nevertheless, the field requires more extensive inquiry and analysis. Additionally, detector-free designs are gaining in popularity. In these designs, the detector is replaced by a general visual encoder, which produces grid features for later cross-modal fusion during vision-language pre-training in an end-to-end manner[35]. Nonetheless, further investigation is necessary to develop a more robust detector-free image captioning model.

The use of image captioning to aid the visually impaired has limited research. Nonetheless, this method could enable a vision assistant suitable for daily use, alerting them to potential hazards and help them comprehend the environment. Unsupervised learning and unpaired settings may fill current gaps, along with the graph-based approach, which is growing in popularity. Using transformers combined with vision-language pre-training may also become a standard practice.

CONCLUSION

This paper provides a comprehensive overview of recent image captioning methods and their features. We categorize the approaches and discuss common problems, datasets, evaluation metrics, and the challenges and future directions in image captioning. Despite the presented solutions, there are still significant challenges in achieving higher quality captions that match human-generated ones. Moreover, evaluating the models' performance remains problematic due to imperfect metrics and limited datasets. However, Vision-Language Pre-Training (VLP) methods and Transformers show significant promise and are likely to be essential components of future image captioning models.

Additionally, we need more research to develop visual assistants that are tailored to meet the unique needs of visually impaired individuals. To make these assistants truly helpful, they must have distinct features that set them apart from other image captioning applications. While advanced models resulting from research may not be the best fit for visual assistants, we still have the potential to create suitable captions for visually impaired individuals. These captions should highlight the most important elements of an image, and provide descriptions of other noticeable and finer details, such as object textures and relative positions. This means that captions designed for visually impaired individuals will necessarily be more detailed and elaborate than those produced through traditional methods and models. We can also modify the process of caption generation, so that users receive a brief, general caption at first, and can then ask for more detailed descriptions based on their own inquiries. Since there are now more visually impaired individuals than ever before, it's important to address these issues and find a solution that is effective and efficient. Valuable research in this area would focus on automatic image captioning with a specific emphasis on designing visual assistants that cater to the needs of visually impaired individuals in a friendly and welcoming way.

REFERENCES

[1] S. Yagcioglu, E. Erdem, A. Erdem, and R. Cakıcı, “A distributed representation based query expansion approach for image captioning,” in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, vol. 10, no. 3115, Beijing, China, July 2015.

[2] H. Fang, S. Gupta, F. Iandola et al., “From captions to visual concepts and back,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, June 2015.

[3] R. Girshick, J. Donahue, D. Trevor, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587, Columbus, OH, USA, June 2014.

[4] C. Zhang, J. C. Platt, and V. Paul, “Multiple instance boosting for object detection,” in Advances in Neural Information Processing Systems 18, pp. 1417–1424, MIT Press, London, UK, 2005.

[5] J. Devlin, H. Cheng, H. Fang, S. Gupta, Li Deng, and X. He, “Language models for image captioning: the quirks and what works,” Computer Science, 2015, http://arxiv.org/abs/1505.01809.

[6] K. Tran, X. He, L. Zhang, and J. Sun, “Rich image captioning in the wild,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 434–441, Las Vegas, NV, USA, June 2016.

[7] P. Razvan, G. Caglar, K. Cho, and B. Yoshua, “How to construct deep recurrent neural networks,” Computer Science, 2014, http://arxiv.org/abs/1312.6026.

[8] T. Mikolov, M. Karafiat, L. Burget, J. “Honza” Cernocky, and S. Khudanpur, “Recurrent neural network based language model,” in Proceedings of the Eleventh Annual Conference of the International Speech Communication Association, DBLP, pp. 1045–1048, Chiba, Japan, September 2010.

[9] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, “Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech,” in Proceedings of the 9th ISCA Speech Synthesis Workshop, pp. 146–152, Sunnyvale, CA, USA, September 2016.

[10] S. O. Arik, M. Chrzanowski, A. Coates, and G. Diamos, “Deep voice: real-time neural text-to-speech,” 2017, http://arxiv.org/ abs/1702.07825.

[11] R. Pascanu, T. Mikolov, and Y. Bengio, “On the difficulty of training recurrent neural networks,” International Conference on Machine Learning, vol. 52, no. 3, pp. 1310–1318, 2012.

[12] W. Zaremba, I. Sutskever, and O. Vinyals, “Recurrent neural network regularization,” 2014, http://arxiv.org/abs/1409.2329.

[13] J. Song, H. Zhang, X. Li, L. Gao, M. Wang, and R. Hong, “Selfsupervised video hashing with hierarchical binary auto-encoder,” IEEE Transactions on Image Processing, vol. 27, no. 7, pp. 3210–3221, 2018.

[14] X. Chen, H. Fang, T.-Yi Lin et al., “Microsoft COCO captions: data collection and evaluation server,” Computer Science, 2015, http://arxiv.org/abs/1504.00325.

[15] 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. International Journal of Computer Vision, 123(1):74–93, May 2017. ISSN 1573-1405. doi:10.1007/s11263-016-0965-7. URL https://doi.org/10.1007/s11263-016-0965-7

[16] M. Hodosh, P. Young, and J. Hockenmaier, “Framing image description as a ranking task: data, models and evaluation metrics,” Journal of Artificial Intelligence Research, vol. 47, pp. 853–899, 2013

[17] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, “Collecting image annotations using Amazon’s Mechanical Turk,” in Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk, pp. 139–147, Los Angeles, CA, USA, June 2010.

[18] 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, 123(1):32–73, May 2017. ISSN 15731405. doi:10.1007/s11263-016-0981-7.

[19] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL ’02, pages 311– 318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. doi:10.3115/1073083.1073135. URL https://doi.org/10.3115/1073083.1073135.

[20] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In Proceedings of the workshop on text summarization branches out (WAS 2004), pages 74–81, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1000.

[21] S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In J. Goldstein, A. Lavie, C.-Y. Lin, and C. Voss, editors, Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, volume 29, pages 65–72, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics. URL https://aclanthology.org/ W05-09.

[22] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In Proceedings of the IEEE conference on computer vision and pattern recognition, volume 07-12-June-2015, pages 4566–4575, Los Alamitos, CA, USA, 2015. IEEE Computer Society. doi:10.1109/CVPR.2015.7299087. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2015.7299087.

[23] O. Vinyals, T. Alexander, S. Bengio, and D. Erhan, “Show and tell: a neural image caption generator,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156–3164, Columbus, OH, USA, June 2014.

[24] K. Xu, J. Ba, K. Ryan et al., “Show, attend and tell: neural image caption generation with visual attention,” in Proceedings of the IEEE Conference on Computer Vision

and Pattern Recognition, pp. 2048–2057, Boston, MA, USA, June 2015. [70] A. Vaswani.

[25] B. Dzmitry, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," Computer Science, 2014, http://arxiv.org/abs/1409.0473.

[26] J. L. Ba, M. Volodymyr, and K. Koray, "Multiple object recognition with visual attention," Computer Science, 2014, http://arxiv.org/abs/1412.7755.

[27] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: semantic propositional image caption evaluation," in Computer Vision—ECCV 2016, vol. 11, pp. 382–398, no. 4, Springer, Cham, Switzerland, 2016.

[28] L. Chen, H. Zhang, J. Xiao et al., "SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6298–6306, Las Vegas, NV, USA, June-July 2016.

[29] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: adaptive attention via a visual sentinel for image captioning," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3242–3250, Las Vegas, NV, USA, June-July 2016.

[30] Y. Li, Y. Pan, T. Yao, and T. Mei. Comprehending and ordering semantics for image captioning. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 17969–17978, Los Alamitos, CA, USA, jun 2022. IEEE Computer Society. doi:10.1109/CVPR52688.2022.01746. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01746.

[31] J. Gu, S. Joty, J. Cai, H. Zhao, X. Yang, and G. Wang. Unpaired image captioning via scene graph alignments. In Proceedings of the IEEE/CVF International Conference on Computer Vision, volume 2019-October, pages 10323–10332, Manhattan, New York, U.S., 2019. IEEE. doi:10.1109/ICCV.2019.01042. URL http://doi.org/10.1109/ICCV.2019.01042.

[32] N. Xu, A.-A. Liu, J. Liu, W. Nie, and Y. Su. Scene graph captioner: Image captioning based on structural visual representation. Journal of Visual Communication and Image Representation, 58:477–485, Jan. 2019. ISSN 1047-3203. doi:https://doi.org/10.1016/j.jvcir.2018.12.027.

[33] X. Yang, K. Tang, H. Zhang, and J. Cai. Auto-encoding scene graphs for image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10685–10694, Manhattan, New York, U.S., 2019. IEEE. doi:10.1109/CVPR.2019.01094. URL http://doi.org/10.1109/CVPR.2019.01094.

[34] K.H. Lee, H. Palangi, X. Chen, H. Hu, and J. Gao. Learning visual relation priors for image-text matching and image captioning with neural scene graph generators. arXiv preprint arXiv:1909.09953, 2019. URL https://arxiv.org/abs/1909.09953.

[35] Z. Fang, J. Wang, X. Hu, L. Liang, Z. Gan, L. Wang, Y. Yang, and Z. Liu. Injecting semantic concepts into end-to-end image captioning. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 17988–17998, Los Alamitos, CA, USA, jun 2022. IEEE Computer Society. doi:10.1109/CVPR52688.2022.01748. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01748.

[36] V. Mnih, N. Heess, and A. Graves, "Recurrent models of visual attention," Advances in Neural Information Processing Systems, vol. 3, pp. 2204–2212, 2014.

[37] K. Xu, J. Ba, K. Ryan et al., "Show, attend and tell: neural image caption generation with visual attention," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2048–2057, Boston, MA, USA, June 2015.

[38] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," in Proceedings of the Neural Information Processing Systems, Long Beach, CA, USA, December 2017.

[39] L. Minh-0ang, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in Proceedings of the Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, September 2015.

[40] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: adaptive attention via a visual sentinel for image captioning," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3242–3250, Las Vegas, NV, USA, June-July 2016.

[41] L. Chen, H. Zhang, J. Xiao et al., "SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6298–6306, Las Vegas, NV, USA, June-July 2016.