Integrated Science

Journal of Integrated
SCIENCE & TECHNOLOGY

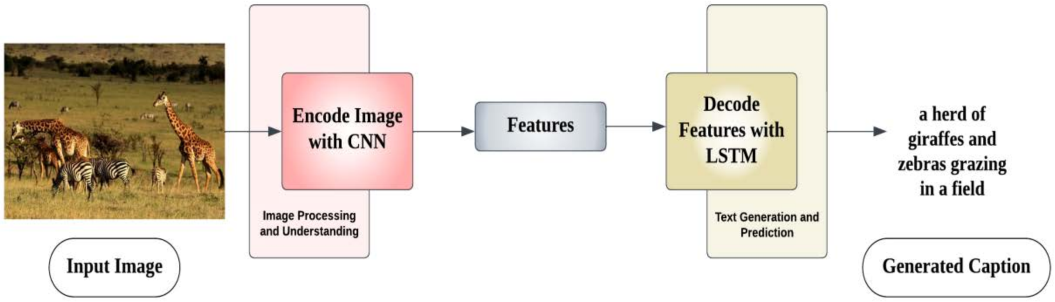
A review of Deep learning image captioning approaches

Yugandhara A. Thakare,[1*] Kishor H. Walse,[2]

[1]PG Department of Computer Science, Sant Gadge Baba Amravati University, Amravati, Maharashtra, India. [2]Sant Bhagwan Baba Kala Mahavidyalaya, Sindkhed Raja, Maharashtra, India.

Received on: 08-Jul-2023, Accepted and Published on: 04-Sept-2023

ABSTRACT


Encode Image with CNN
Image Processing and Understanding
Features
Decode Features with LSTM
Text Generation and Prediction
a herd of giraffes and zebras grazing in a field
Input Image
Generated Caption

In today's information-driven world, images have become a prevalent and influential means of communication and artistic expression. While humans effortlessly understand visual scenes and describe them in nuanced language, replicating this ability in machines has been a significant hurdle. Image captioning, a burgeoning field at the intersection of computer vision and natural language processing (NLP), aims to overcome this challenge by developing sophisticated algorithms and models that can intelligently interpret visual data and generate accurate, contextually relevant, and human-like textual descriptions for images. This survey paper presents a systematic examination of deep learning approaches in image captioning, offering a detailed taxonomy for each method category. It extensively covers widely-used datasets and evaluation metrics designed to assess image captioning model performance. The discussion emphasizes challenges encountered in the field along with highlighting the current state-of-the-art technologies.

Keywords: Image captioning, Computer vision, Natural language processing (NLP), Attention mechanism, Deep learning

INTRODUCTION

In the era of visual information abundance, images have emerged as a dominant medium for communication, storytelling, and artistic expression. Humans possess an innate ability to effortlessly perceive the content of visual scenes and eloquently describe them with rich linguistic context. However, endowing machines with the capacity to understand and generate meaningful textual descriptions for images has been a persistent challenge. The captivating research field of image captioning, at the convergence of computer vision and natural language processing (NLP), seeks to overcome this hurdle by developing sophisticated algorithms and models capable of intelligently interpreting visual content and translating it into coherent and contextually relevant captions.

Image captioning has witnessed significant advancements with the advent of deep learning and the integration of powerful neural network architectures. Convolutional Neural Networks (CNNs) have revolutionized image feature extraction, enabling models to capture intricate visual patterns and representations. Meanwhile, Recurrent Neural Networks (RNNs) and Transformers have facilitated more robust and expressive language modeling, leading to the generation of fluent and contextually coherent textual descriptions.

As the field of image captioning continues to evolve, a primary focus lies on achieving generalization across diverse and complex visual scenes. Early approaches often struggled to handle variations in images, leading to limited applicability and overfitting on specific domains.[1] To address this, researchers have embraced large-scale datasets, such as MSCOCO and Visual Genome, which encompass diverse visual concepts and contexts. These datasets, paired with advanced evaluation metrics like BLEU, METEOR, and CIDEr, serve as valuable benchmarks for assessing model

*Ms. Yugandhara A. Thakare, Research Scholar, PG Department of Computer Science, SGBAU, Amravati, Maharashtra, India.
Email: yugathakare@gmail.com

Cite as: *J. Integr. Sci. Technol., 2024, 12(1), 712.*
URN:NBN:sciencein.jist.2024.v12.712

ISSN: 2321-4635
http://pubs.thesciencein.org/jist

performance and pushing the boundaries of image captioning. In Figure 1, the taxonomy presented in this research is illustrated.

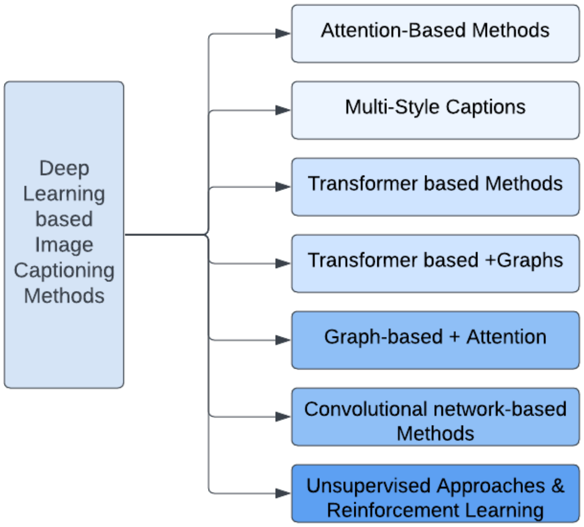

Deep Learning based Image Captioning Methods
Attention-Based Methods
Multi-Style Captions
Transformer based Methods
Transformer based +Graphs
Graph-based + Attention
Convolutional network-based Methods
Unsupervised Approaches & Reinforcement Learning

Figure 1. The classification of the image captioning techniques discussed in this survey article

Deep Learning-Based Image Captioning

2.1 **Attention-Based Methods**

Attention-based image captioning methods leverage attention mechanisms to improve the quality and relevance of generated captions. These mechanisms enable the model to focus on specific regions of the image while generating each word of the caption, allowing it to align visual features with the corresponding linguistic context better. This results in more accurate, descriptive, and contextually relevant captions. This dynamic attention mimics the human visual system, improving caption quality significantly. In the broader context, attention mechanisms have reshaped image captioning by enhancing the model's ability to recognize and describe fine-grained details in images, contributing to more contextually relevant and descriptive captions.

Show, Attend and Tell (SAT) - 2015: The SAT system employs both Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) for image feature extraction and language modeling, respectively. What sets it apart is the incorporation of attention mechanisms. This enables the model to focus on specific regions of the image while generating each word of the caption. The attention weights are learned based on how relevant different image regions are to the current word being produced. Figure 2 depicts a block diagram illustrating the framework of attention-based method.[2]

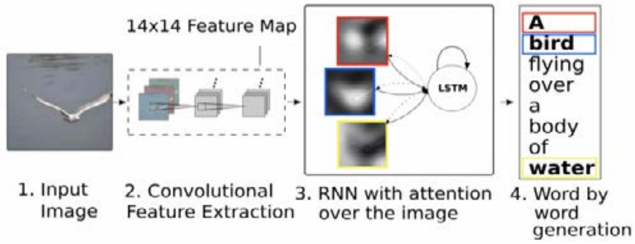

14x14 Feature Map
LSTM
A
bird
flying
over
a
body
of
water
1. Input Image
2. Convolutional Feature Extraction
3. RNN with attention over the image
4. Word by word generation

Figure 2. The general workflow of attention-based method [2].

Adaptive Attention (AdaAtt) - 2017: The AdaAtt model extends the SAT approach by introducing adaptive attention mechanisms. It uses a combination of soft and hard attention, allowing the model to dynamically switch between attending to image regions softly and attending to a single hard region. This adaptive attention mechanism improves the model's flexibility in handling different types of images and improves caption quality.[3]

Bottom-Up and Top-Down Attention - 2018: This method integrates bottom-up region proposals, obtained from object detectors, with top-down attention, enabling the model to selectively focus on significant image regions in a more informative manner. By leveraging both the low-level visual features and high-level context, the model produces captions that are more precise and detailed.[4]

Self-critical Sequence Training (SCST)-2017SCST introduces a reinforcement learning technique to fine-tune the captioning model. Instead of relying on the conventional cross-entropy loss, SCST adopts a reward-based approach, where the model is trained to optimize a reward function (such as CIDEr or BLEU). This reward function evaluates the quality of the generated captions. This method significantly improves the model's performance and caption diversity.[5]

AoANet (Attention on Attention Network) - 2019: AoANet proposes a novel "attention on attention" mechanism, where the model learns to attend to different parts of the image as well as different parts of the language context simultaneously. This multi-level attention approach improves the alignment between visual content and textual context, leading to more accurate and contextually relevant captions.[6]

M2 Transformer - 2020: The M2 Transformer model utilizes a multi-modal transformer, which integrates both image and text modalities seamlessly. The transformer-based architecture captures long-range dependencies more effectively and facilitates better context understanding for improved caption generation.[7]

Look, Imagine and Match 2020: This work focuses on cross-modal retrieval but introduces a technique called "looking and imagining," where the model attends to the input image and simultaneously imagines additional information to improve the relevance of the generated captions. This approach suggests a way to enhance the captioning process by combining image understanding with creative imagination, leading to more contextually relevant and informative captions.[8]

Overall, these reports contribute to the ongoing research in attention-based image captioning by introducing innovative techniques to improve the quality, relevance, and context-awareness of generated captions. By incorporating attention mechanisms and adaptive fusion techniques, these methods advance the state-of-the-art in image captioning and offer promising avenues for generating more accurate and contextually meaningful captions in various applications. Researchers continue to build upon these findings to further refine and enhance the capabilities of attention-based image captioning models. Comparison of attention-based methods is shown in Table 1 based on approach and a drawback parameter. Table 2 compares attention-based approaches using standard evaluation metrics.

Table 1. Comparison of Attention-Based Methods based on methodology and drawback parameter.

Title of Paper	Methodology	Drawback
Show, Attend and Tell (SAT) - 2015	SAT employs a blend of Convolutional Neural Networks (CNNs) for extracting image features and Long Short-Term Memory (LSTM) for language modeling. The significant advancement lies in incorporating attention mechanisms, enabling the model to focus selectively on distinct image regions when generating each caption word. The attention weights are learned based on the relevance of specific image regions to the word being generated at that moment.	One drawback of SAT could be its complexity, as the introduction of attention mechanisms may increase the computational cost and training time. Additionally, it might be challenging to handle very large images effectively, impacting the model's scalability.
Adaptive Attention (AdaAtt) - 2017	The AdaAtt model extends the SAT approach by introducing adaptive attention mechanisms. It uses a combination of soft and hard attention, allowing the model to dynamically switch between attending to image regions softly and attending to a single hard region. This adaptive attention mechanism improves the model's flexibility in handling different types of images and improves caption quality.	A potential drawback of AdaAtt could be the increased complexity compared to SAT due to the introduction of both soft and hard attention mechanisms. This might result in higher memory requirements and longer training times. Moreover, the performance gains may vary depending on the specific image dataset and task at hand.
Bottom-Up and Top-Down Attention - 2018	This method merges bottom-up region proposals, produced using object detectors, with top-down attention, enabling the model to pay closer attention to pertinent image regions in a more informative manner. By leveraging both the low-level visual characteristics and high-level context, the model generates captions that are more precise and elaborate.	One limitation of Bottom-Up and Top-Down Attention is that it relies on object detectors to generate region proposals, which might not always accurately capture all relevant image regions. Additionally, the reliance on pre-generated region proposals might hinder the model's ability to adapt to varying image characteristics.
Self-critical Sequence Training (SCST) - 2017	SCST introduces reinforcement learning to refine the captioning model. Instead of relying on the conventional cross-entropy loss, SCST adopts a reward-based strategy. In this approach, the model is trained to maximize a reward function (e.g., CIDEr or BLEU) that assesses the quality of the generated captions. As a result, this technique notably enhances the model's performance and fosters a greater diversity in the captions it produces.	A drawback of SCST is that it requires additional computation during training due to the use of reinforcement learning, which may lead to longer training times. Furthermore, defining an appropriate reward function for reinforcement learning can be challenging, potentially impacting the quality of the generated captions.
AoANet (Attention on Attention Network) - 2019	AoANet proposes a novel "attention on attention" mechanism, where the model learns to attend to different parts of the image as well as different parts of the language context simultaneously. This multi-level attention approach improves the alignment between visual content and textual context, leading to more accurate and contextually relevant captions.	One limitation of AoANet could be the increased complexity introduced by the multi-level attention approach, which might lead to higher memory and computational requirements. Additionally, interpreting the attention patterns of the "attention on attention" mechanism may be more challenging than standard attention mechanisms.
M2 Transformer - 2020	The M2 Transformer model utilizes a multi-modal transformer, which integrates both image and text modalities seamlessly. The transformer-based architecture captures long-range dependencies more effectively and facilitates better context understanding for improved caption generation.	A potential drawback of M2 Transformer could be the increased computational complexity compared to traditional LSTM-based approaches, which may result in higher training times and resource requirements. Additionally, the effectiveness of capturing long-range dependencies may vary depending on the complexity of the image-text relationships in different datasets.
Look, Imagine and Match: Improving Textual-Visual Retrieval...	The primary emphasis of this paper lies in cross-modal retrieval. However, it introduces a method termed "looking and imagining," where the model not only attends to the input image but also conjures up additional information concurrently. This technique aims to enhance the relevance of the generated captions. This approach suggests a way to enhance the captioning process by combining image understanding with creative imagination, leading to more contextually relevant and informative captions.	Since this paper primarily emphasizes cross-modal retrieval, one potential limitation could be the relatively limited evaluation of the model's performance on the image captioning task alone compared to specialized captioning models. Moreover, the efficacy of the "looking and imagining" technique could be subject to variations based on factors such as the complexity of the image and the quality of the additional information generated by the model's imagination.

Table 2. Comparison of Attention-Based Methods based on standard evaluation metrics

Ref.	Dataset	B@1	B@2	B@3	B@4	METEOR	CIDEr	ROUGE -L
[2]	COCO	71.8	50.4	35.7	25.0	23.04	-	-
[3]	COCO	0.742	0.580	0.439	0.332	0.266	1.085	0.549
[4]	COCO	80.2	64.1	49.1	36.9	27.6	117.9	57.1
[5]	COCO	78.1	61.9	47.0	35.2	27.0	114.7	56.3
[6]	COCO	81.0	65.8	51.4	39.4	29.1	126.9	58.9
[7]	COCO	81.6	66.4	51.8	39.7	29.4	129.3	59.2
[8]	COCO	57.1	36.9	23.3	14.9		61.1	

2.2 **Generating Multi-Style Captions**

Multi-style captioning methods have gained importance due to the need for versatile and adaptable image captioning systems. These methods enable the generation of captions with varying tones, emotions, or languages to cater to diverse user preferences and applications. In a world where content personalization and creative content generation are essential, multi-style captioning techniques play a vital role by offering a spectrum of captioning styles for different contexts.

The studies that have been discussed thus far produce neutral captions. These automated captions typically include factual information about the contents of the images. Humans converse with one another on a daily basis using a variety of speech patterns and tones. Humorous, angry, and poetic styles and tones are among them. By incorporating these styles, captions will seem better and encourage more human interaction. Applications like photo-sharing and chatbots can also make use of stylized captions. In addition to including photos and the corresponding captions, Shuster et al.'s dataset[9] also includes features for tone and style. In this study, a novel structure called TransResNet is presented, employing an encoder-decoder architecture to map images, captions, and their associated personality attributes into a shared space. Two types of models were considered: retrieval models and generative models. The generative model utilizes the TransResNet structure to generate captions word by word, whereas the retrieval model considers any caption in the dataset as a possible candidate response. The retrieval model yielded superior results compared to the generative model.

Guo et al.[10] have presented a structure made up of five components for caption generating in various styles. An encoder for plain images is the first module. The caption generator module follows, which generates a text based on a predetermined style. The following module is a caption discriminator designed to distinguish genuine sentences from fake ones. It is trained adversarially, meaning it encourages the learning process to generate more compelling captions that resemble human language better. The style of the resulting caption is then determined using a style discriminator module. Another module dubbed "The Back-Translation Module is employed since neutral and styled captions share some content similarities. This module converts a decorative caption into a neutral one. Multilingual neural machine translation (NMT) is used to carry out this procedure, and neutral captions are considered the output while styled captions are considered the input. Figure 3 illustrates the overall process of multi style captions.

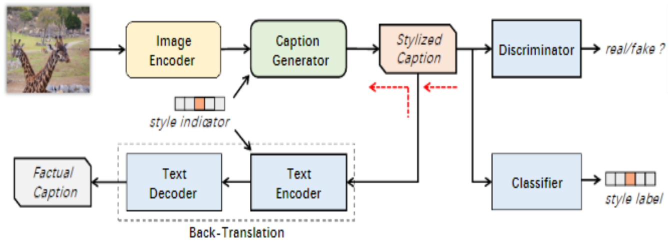


Figure 3. The general workflow of multi style captions [10].

2.3 **Transformer-based**

Transformer architectures have become widely popular in diverse natural language processing tasks, including image captioning, mainly because of their remarkable capability to effectively capture long-range dependencies. Transformer-based models have demonstrated remarkable success in natural language processing tasks. In image captioning, they offer a robust framework for modeling complex relationships between visual and textual information, surpassing earlier architectures. Transformer-based approaches are highly relevant in the research landscape because they consistently achieve state-of-the-art results, pushing the boundaries of image captioning quality and accuracy. Below are a few noteworthy works in the field of Transformer-based image captioning:

Unified Vision-Language Pre-training for Image Captioning and VQA (2019) introduces a cohesive vision-language pre-training framework that utilizes Transformers to acquire shared representations for both image captioning and visual question answering (VQA) tasks. Pre-training on large-scale image-text data enables better performance on downstream tasks like image captioning.[11]

UNITER: Universal Image-text Representation Learning is a pre-training method that employs the Transformer architecture to acquire a universal joint image-text representation. The model is initially trained on extensive image-text data and subsequently fine-tuned for multiple downstream tasks, including image captioning.[12]

ImageBERT is a cross-modal pre-training approach that makes use of vast amounts of weakly supervised image-text data. This Transformer-based model is first pre-trained on this data and later fine-tuned for particular downstream tasks, such as image captioning.[13]

Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks (2021) introduces a pre-training technique for vision-language tasks, which includes image captioning, by aligning the semantics of objects in images with their corresponding words. The approach leverages the Transformer-based architecture to improve the learning of image-text representations.[14]

Figure 4 depicts the general workflow of transformer-based captions. Comparison of Transformer-Based methods is shown in Table 3 based on approach and a drawback parameter.

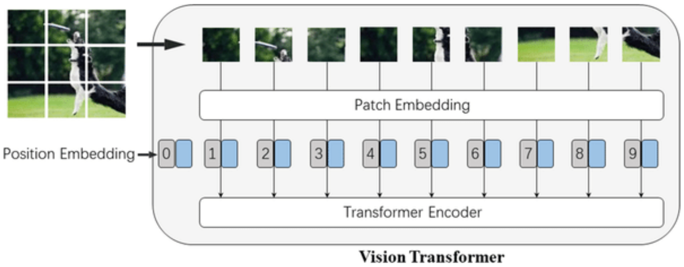


Figure 4. The general workflow of transformer based captions[15]

Overall, the literature review demonstrates the wide range of applications and benefits of Transformer-based methods in image captioning and vision-language tasks. These approaches have pushed the boundaries of image captioning performance by effectively capturing context and generating more accurate, diverse, and contextually relevant captions. As research in this area continues to evolve, we can expect further advancements and innovations in Transformer-based image captioning, contributing

Table 3. Comparison of Transformer-Based Methods based on methodology and drawback parameter.

Title of Paper	Methodology	Drawback
Unified Vision-Language Pre-training for Image Captioning and VQA (2019)	The paper proposes a unified vision-language pre-training framework that leverages Transformers to learn joint representations for image captioning and visual question answering (VQA). Pre-training on large-scale image-text data enables better performance on downstream tasks like image captioning.	A limitation of the unified pre-training approach could be the potential bias introduced by the large-scale image-text data used during pre-training. The model's performance on specific downstream tasks might be influenced by the distribution and quality of the pre-training data, which may not always perfectly match the target task's distribution. Careful selection of pre-training data is essential to ensure generalization to various image captioning tasks.
UNITER: Universal Image-text Representation Learning (2020)	UNITER is a pre-training approach that learns a universal joint image-text representation using the Transformer architecture. The model is pre-trained on large-scale image-text data and then fine-tuned for various downstream tasks, including image captioning.	One potential drawback of UNITER could be the substantial computational resources required for pre-training on large-scale datasets. The massive number of parameters in the Transformer can make pre-training computationally expensive and time-consuming. Additionally, while the pre-trained representations are general, fine-tuning on specific image captioning tasks may still require careful parameter tuning to achieve optimal performance.
ImageBERT: Cross-Modal Pre-training with Large-scale Weak-supervised Image-text Data (2020)	ImageBERT is a cross-modal pre-training method that utilizes large-scale weakly supervised image-text data. The Transformer-based model is pre-trained on this data and then fine-tuned for specific downstream tasks, such as image captioning.	A potential limitation of ImageBERT could be the reliance on weakly supervised data for pre-training. Weak supervision might lead to noisier and less informative pre-training data compared to fully supervised approaches, potentially affecting the quality and generalization of the pre-trained representations. The performance gains from weakly supervised pre-training might be more limited compared to fully supervised alternatives.
Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks (2021)	Oscar proposes a pre-training method for vision-language tasks, including image captioning, by aligning object semantics in images with their corresponding words. It utilizes the Transformer-based architecture to learn better image-text representations.	A limitation of Oscar could be the interpretability of the learned object semantics alignment. While aligning object semantics can be beneficial for certain tasks, the model's ability to explicitly associate objects with words might be limited by the complexity of the pre-training data and the challenge of defining precise object-word relationships in diverse visual contexts. Careful design and analysis are required to ensure the alignment captures meaningful semantics.

to more sophisticated and contextually-aware image understanding systems.

2.4 Transformer-based +Graphs

The integration of graph structures with Transformers allows for better modeling of relationships and contextual information, leading to improved performance in generating accurate and contextually relevant image captions. Integrating Transformer models with graph structures facilitates capturing intricate relationships within images, such as object interactions and hierarchies. This enhances the model's comprehension of visual content and context. In contexts where understanding visual scenes is essential, like autonomous driving or medical imaging, Transformer-based models combined with graphs contribute significantly to generating accurate and contextually rich image captions. Below are some notable works in this area:

Although study entitled "Graphical Contrastive Losses for Scene Graph Parsing" (CVPR 2019) is not focused on image captioning, this study presents graphical contrastive losses for scene graph parsing. The novel loss function aligns graph representations with contrastive samples, making it adaptable to graph-based image captioning tasks.[16]

Similarly, Meshed-Memory Transformer for Image Captioning (ICCV 2019) introduces a meshed-memory Transformer for image captioning. The model utilizes graph-like connections between word tokens in the caption, enabling the generation of long and coherent captions.[17]

Yu Zhang et.al.[18] reports an advanced approach for generating descriptive captions for images. The authors blend the strengths of transformer-based models, recognized for their efficacy in natural language processing tasks, with the incorporation of knowledge graphs. By leveraging knowledge graphs, the model gains access to external information and context, leading to the generation of more informed and contextually relevant captions. The paper likely delves into the model architecture, training procedure, and evaluation outcomes to demonstrate the effectiveness of their proposed approach.[18]

The integration of graph structures with Transformers in image captioning has shown promising results, enabling better modeling of relationships and contextual information in complex visual scenes. These methods open up new possibilities for more accurate and contextually relevant image captioning, and they also pave the way for further advancements in graph-based deep learning techniques in the computer vision domain.

In conclusion, the application of graph-based techniques, such as graph convolutional networks and graphical contrastive losses, along with the integration of graph structures with visual Transformers, shows great potential in advancing image captioning. These methods provide a deeper understanding of visual relationships and contextual information, leading to more accurate, diverse, and contextually relevant image captions. As the research in this area continues to evolve, we can expect further advancements and innovations in graph-based image captioning techniques, contributing to more sophisticated and effective image understanding systems. Comparison of Transformer-Based + Graphs methods is shown in Table 4 based on approach and a drawback parameter.

Table 4. Comparison of Transformer-Based + Graphs Methods based on methodology and drawback parameter.

Title of Paper	Methodology	Drawback
Graphical Contrastive Losses for Scene Graph Parsing (CVPR 2019)	Although not focused on image captioning, this work introduces graphical contrastive losses for scene graph parsing. The proposed loss function aligns graph representations with contrastive samples, which can be adapted to graph-based image captioning tasks.	As this work is not specifically designed for image captioning, its adaptation to the captioning task may require further research and experimentation. The performance and effectiveness of the graphical contrastive losses for image captioning might differ from their use in scene graph parsing. The selection and design of appropriate contrastive samples for image captioning may also impact the model's performance.
Meshed-Memory Transformer for Image Captioning (ICCV 2019)	The paper introduces a meshed-memory Transformer for image captioning. The model utilizes graph-like connections between word tokens in the caption, enabling the generation of long and coherent captions.	One potential drawback of the meshed-memory Transformer could be the increased memory consumption and computational cost due to the graph-like connections between word tokens. Additionally, the quality of generated captions may depend on the effectiveness of modeling long-range dependencies through the graph connections, which could vary based on the specific captioning task and dataset.

2.5 **Graph-based + Attention**

Graph-based attention methods for image captioning have shown promising results in enhancing the contextual understanding of images and generating more accurate and contextually relevant captions. These approaches leverage both graph structures and attention mechanisms to capture semantic relationships between objects in the image and effectively focus on important image regions during caption generation. Graph-based methods represent visual scenes with structured graph representations, enabling the modeling of complex relationships. Combining these with attention mechanisms enhances the model's capacity to attend to relevant regions and connections in the graph. This approach is highly relevant when fine-grained descriptions, spatial reasoning, and object interactions need to be captured accurately, as in scientific image analysis or robotics applications.

W. Ziang et.al.[19] reported an object-aware multi-attention network for image captioning. The model incorporates graph structures to represent objects and applies multi-level attention mechanisms to focus on different image regions during caption generation.[19]

In order to take into account the relationships between image elements as well as the attention-based encoder-decoder architecture, Yao et al.[20] combine Graph Convolutional Networks[21] with LSTMs.[22] The image encoder has combined spatial and semantic linkages to enhance image representations, and understanding the relationships has been thought of as a classification challenge. Region suggestions have been made using faster R-CNN.[23] To capture the geographic and semantic relationships within image contents, two distinct graphs, namely spatial and semantic graphs, are constructed. The regions identified in the images serve as nodes in these graphs, while the relationships between these regions are represented as the graph's edges. In the semantic graph, the edges signify semantic relations, while in the spatial graph, the edges denote spatial relations. The semantic graph is trained using the Visual Genome dataset,[24] while the image representation is carried out using Graph Convolutional Networks,[21] which incorporate the semantic and spatial relations derived from their corresponding graphs. The caption sentences are then generated using an LSTM[22] decoder, which takes into account the augmented region representations and their associated semantic and spatial links. During inference, at each time step, the outputs from the two decoders (spatial and semantic) are combined using a linear weight sum, and the word with the highest probability is selected. This process effectively integrates the output of both spatial and semantic decoders for generating the final captions.

In conclusion, graph-based attention methods for image captioning have shown significant improvements in capturing semantic relationships between objects and focusing on important image regions during caption generation. By combining graph structures with attention mechanisms, these approaches enhance the contextual understanding of images and produce more accurate and coherent captions. Graph-based attention methods open up new possibilities for further research and innovation in image captioning, contributing to the development of more sophisticated and effective image understanding systems.

2.6 **Convolutional network-based**

Convolutional neural network (CNN)-based image captioning methods have been widely explored in the literature to generate captions for images. These approaches typically use CNNs for image feature extraction and combine them with recurrent neural networks (RNNs) or transformers for language modeling. Convolutional neural networks (CNNs) have long been the cornerstone of image feature extraction. In image captioning, they play a pivotal role in effectively encoding visual information and extracting meaningful features. CNN-based methods remain essential in the field, as they provide the foundational visual features required for generating captions, making them a crucial component in most image captioning pipelines.

'Show and Tell' by O. Vinyals et.al.[25] introduces an early and influential CNN-RNN-based image captioning model. The system utilizes a Convolutional Neural Network (CNN) for extracting image features and a Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM) to sequentially generate captions.[25]

NeurIPS by J. Mao et.al.[26] also introduces a CNN-RNN-based image captioning model with visual attention. The model attends to different image regions while generating each word of the caption to improve caption quality.[26] J. Lu et.al.[27] reported an adaptive attention mechanism that employs a visual sentinel to determine whether to softly attend to various image regions or concentrate on a single, specific region. This flexibility improves caption quality for different types of images.[27]

These works represent a snapshot of the early developments and advancements in CNN-based image captioning methods. Since their publication, the field has evolved, and more sophisticated architectures and techniques have been proposed. CNN-based image captioning remains an active and dynamic research field, striving to produce more precise, contextually relevant, and varied captions for a diverse set of images. The comparison of Convolutional network-based approaches is provided in Table 5 based on approach and a drawback parameter.

Table 5. Comparison of Convolutional network-based method on methodology and drawback parameter.

Title of Paper	Methodology	Conclusion
Show and Tell: A Neural Image Caption Generator (CVPR 2015)	The paper introduces an early and influential CNN-RNN-based image captioning model. It uses a CNN to extract image features and an RNN (LSTM) to generate captions sequentially.	The "Show and Tell" model showcases the effectiveness of using deep learning architectures, specifically CNNs and LSTMs, for image captioning tasks. The sequential generation of captions using an RNN allows the model to capture the temporal dependencies in the language. The approach establishes a strong foundation for subsequent research in image captioning.
Neural Image Captioning with Visual Attention (NeurIPS 2015)	This paper also introduces a CNN-RNN-based image captioning model with visual attention. The model attends to different image regions while generating each word of the caption to improve caption quality.	By incorporating visual attention in the image captioning model, this work achieves improved captioning performance by focusing on relevant image regions. The attention mechanism allows the model to capture fine-grained details and relationships between objects in the image, leading to more informative and contextually relevant captions. This approach showcases the significance of integrating visual attention in CNN-RNN-based captioning models to enhance the overall caption generation process.
Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning (CVPR 2017)	The paper introduces an adaptive attention mechanism that uses a visual sentinel to decide when to attend to different image regions softly or focus on a single hard region. This flexibility improves caption quality for different types of images.	The adaptive attention mechanism proposed in this work enhances the model's ability to handle different image characteristics effectively. By incorporating a visual sentinel, the model can dynamically switch between soft and hard attention, adapting to the specific content of the image. This adaptability results in more contextually relevant and accurate captions. The method highlights the significance of adaptive attention mechanisms in image captioning, enabling the model to enhance its attention based on the context of the image.

2.7 **Unsupervised Approaches & Reinforcement Learning**

Unsupervised approaches and reinforcement learning have been explored in the literature to address the challenges of generating image captions without relying on paired image-caption training data. These approaches aim to utilize extensive image datasets and reinforcement learning techniques to enhance the quality and diversity of the generated captions. Unsupervised image captioning explores methods for generating captions without relying on paired image-text data. This approach is significant for making image captioning cost-effective, adaptable, and applicable to various domains. Unsupervised approaches hold promise in scenarios where labeled data is scarce or expensive to obtain, potentially unlocking new avenues for image captioning in diverse contexts. Reinforcement learning techniques introduce a feedback loop, optimizing captions based on quality metrics. This iterative approach helps models generate more coherent and contextually appropriate captions. Reinforcement learning is crucial for fine-tuning and improving caption generation, pushing the performance of image captioning models to higher levels and aligning them more closely with human-generated captions.

t. Chen et.al.[28] presents a recurrent generative adversarial network (GAN) designed for language generation. The model is trained in an unsupervised manner using reinforcement learning, without requiring pre-training. This showcases its capability to generate captions without the need for paired data.[28]

A. Creswell et.al.[29] reported an unsupervised approach to image captioning using reinforcement learning with no paired data. The model is trained to optimize a reward function that assesses the quality of the generated captions. Remarkably, it achieves competitive results without the need for supervised training.[29] Figure 5 depicts the general workflow of unsupervised method.

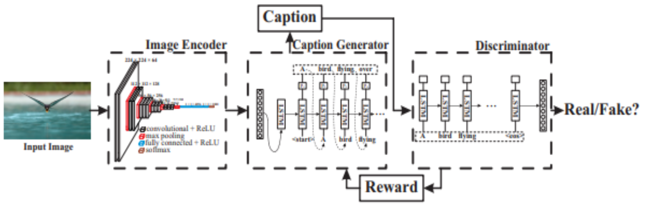


Figure 5: The general workflow of unsupervised method. [39]

These literature reports represent a subset of the research on unsupervised approaches and reinforcement learning for image captioning. They demonstrate the potential of leveraging large-scale image datasets and reinforcement learning to generate high-quality image captions without the need for paired image-caption training data. The field continues to evolve, and further advancements are expected in the quest for more accurate, diverse, and unsupervised image captioning methods. Comparison of Unsupervised Approaches & Reinforcement Learning methods is shown in Table 6 based on approach and a drawback parameter.

ETHICAL CONSIDERATIONS AND POTENTIAL BIASES ASSOCIATED WITH IMAGE CAPTIONING

When using huge datasets and pre-trained models, ethical issues and potential biases are crucial aspects of image captioning. Large-scale data sets used for training may unintentionally add biases, reflecting societal prejudices or skewed representations, resulting in the generation of captions that maintain stereotypes or reflect cultural insensitivity. Additionally, pre-trained models may have difficulty handling images that are outside the scope of their training, running the risk of underrepresentation or misinterpretation. Users who encounter offensive or unsuitable captions may lose faith in these applications, and accidentally disclosing important information may give rise to privacy

Table 6. Comparison of Unsupervised Approaches & Reinforcement Learning method based on methodology and drawback parameter.

Title of Paper	Methodology	Conclusion
Language Generation with Recurrent Generative Adversarial Networks without Pre-training (ICLR 2017)	The paper introduces a recurrent generative adversarial network (GAN) for language generation. The model is trained in an unsupervised manner using reinforcement learning without pre-training, demonstrating its potential for generating captions without paired data.	The unsupervised language generation with recurrent GANs highlights the ability to generate captions without the need for pre-training on paired data. The approach demonstrates the effectiveness of reinforcement learning in training a model to produce high-quality captions based on rewards. This work opens up possibilities for unsupervised approaches to language generation and image captioning, reducing the dependency on annotated data.
Unsupervised Image Captioning (ECCV 2018)	The paper proposes an unsupervised approach to image captioning using reinforcement learning with no paired data. The model is trained to maximize a reward function that evaluates the quality of generated captions, achieving competitive results without supervised training.	The unsupervised image captioning approach demonstrates competitive performance in generating captions without the need for paired data. By optimizing the model based on a reward function, the method can produce high-quality captions in an unsupervised manner. This work showcases the potential of reinforcement learning in training captioning models and its applicability in unsupervised settings, reducing the dependency on annotated data and advancing the field of unsupervised image captioning.

problems. In order to effectively address these ethical concerns and biases, it is important to establish ethical norms and responsible practices in image captioning development. Ensuring accountability and transparency can be difficult when working with complicated models.

Datasets utilized in recent works

Here, we provide a detailed analysis of the datasets that have been commonly used in recent research works. In comparison to object detection, the datasets for the image captioning task are still quite limited. Given the task's growing significance, creating richer datasets may be essential to the task's growth and advancement. Details about the most popular datasets are shown in table 7.

3.1 The MS COCO Dataset[30] is an extensive collection used for tasks like object detection, image segmentation, and image captioning. It boasts numerous attributes, including image segmentation, a massive collection of 328,000 images, a wide variety of 91 object classes, and the unique characteristic of having five captions associated with each image.

3.2 The Flickr30K Dataset[31] is specifically curated for automatic image captioning and grounded language understanding tasks. It comprises 31,000 images gathered from Flickr and is accompanied by 158 thousand human-written captions. The dataset includes detectors for common objects, a color classifier, and a preference for larger objects. Notably, it lacks predefined split settings for training, testing, and evaluation, granting researchers the flexibility to use any desired splits according to their needs.

3.3 The Visual Genome Dataset[24] is distinct from other datasets as it provides individual captions for each region of an image. This extensive dataset comprises seven elements, including region descriptions, attributes, relationships, region graphs, scene graphs, and question-answer pairs. It encompasses more than 108 thousand images, with each image containing an average of 35 objects, 26 attributes, and 21 pairwise relationships between the objects.

3.4 The Conceptual Captions Dataset[32] contains approximately 3.3 million image-caption pairs. The images were initially sourced from the Internet, accompanied by their corresponding "alt-text." The dataset was carefully curated to extract appropriate captions that precisely describe the image contents. It is divided into training and evaluation sets, with 3,318,333 image-caption pairs in the training set and 15,840 image-caption pairs in the evaluation set.

3.5 The FlickrStyle10k Dataset[33] is composed of 10 thousand images, each accompanied by captions in various styles. The training set contains 7 thousand images, whereas the testing and evaluation sets comprise 2 thousand and 1 thousand images, respectively. Each image in the dataset is paired with captions written in various styles, including poetic, humorous, and neutral (factual) styles.

Table 7. Information of most common Datasets

Datasets Name	Total Images	Objects per Image	Object Classes	Captions per Image
MS COCO	330,000	7.57	91	5
Flickr30k	31,000	-	-	5
Visual Genome	108,077	36.17	80,138	5
Conceptual Captions	3.3M	-	-	1
FlickrStyle10K	10,000	-	-	2

Evaluation Metrics for Image Captioning Methods

BLEU (Bilingual Evaluation Understudy): is a metric used to assess the similarity between the generated caption and reference captions by calculating the n-gram precision. It quantifies how closely the generated caption aligns with the reference captions.[34]

4.2 METEOR (Metric for Evaluation of Translation with Explicit ORdering): METEOR is another metric based on n-gram precision but also considers recall and aligns words based on their meanings. It accounts for synonyms and paraphrases, making it suitable for evaluating diverse captions.[35]

4.3 CIDEr (Consensus-based Image Description Evaluation) :calculates the agreement between the generated caption and reference captions. It evaluates the similarity of the generated caption with all the reference captions, taking into account the diversity of human-provided captions.[36]

4.4 ROUGE (Recall-Oriented Understudy for Gisting Evaluation): is a metric originally developed for assessing text summarization tasks, but it has been adapted for image captioning evaluations. It quantifies the overlap between the generated caption and the reference captions by considering n-grams and longest common subsequence.[37]

4.5 SPICE (Semantic Propositional Image Caption Evaluation): is a metric used to assess the quality of generated captions by calculating the semantic similarity between the generated caption and reference captions based on semantic propositions.[38]

4.6 ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation - Longest Common Subsequence) is a variant of ROUGE designed to focus on the longest common subsequence between the generated caption and reference captions. It is widely used in image captioning evaluations.[37]

Challenges in Image Captioning Task

Numerous approaches and solutions have been proposed to tackle the challenges of image captioning. However, several obstacles and unresolved issues persist, some of which are listed below:

Ambiguity: Images frequently possess multiple valid interpretations, which can result in ambiguity during caption generation. Selecting the most suitable and contextually relevant caption for an image presents a challenging task.

Diversity: Generating diverse and creative captions is important to avoid repetitive and generic descriptions. Ensuring diversity while maintaining the accuracy and coherence of the captions is a challenge.

Fine-Grained Understanding: Accurately capturing fine-grained details and intricate relationships between objects in complex images remains a challenging task. The model must comprehend subtle visual cues to generate precise and descriptive captions.

Out-of-Distribution Images:Image captioning models may encounter difficulties when confronted with images that lie outside their training distribution. Adapting the model to handle unseen or out-of-distribution images poses a challenging task.

Handling Long Captions: Generating lengthy and coherent captions for images with complex scenes can be difficult. Ensuring that the model maintains consistency and context throughout the caption is a challenge.

Efficient Inference: Deploying image captioning models in real-time applications requires efficient inference. Optimizing the models for low-latency and resource-efficient execution is a challenge.

Future Directions in Image Captioning

Multi-modal Understanding: Advancing image captioning models to better integrate visual and textual information for more comprehensive understanding. Incorporating contextual information and world knowledge to enhance the model's caption generation capabilities.

Reinforcement Learning: Further exploring the use of reinforcement learning for image captioning to optimize evaluation metrics directly and encourage diverse and creative caption generation.

Pre-training and Transfer Learning: Leveraging pre-trained models from large-scale image-text datasets can enhance the performance of image captioning models. Transfer learning approaches can help to generalize across domains and handle out-of-distribution images better.

Multilingual Image Captioning: Extending image captioning models to generate captions in multiple languages, enabling cross-lingual understanding and accessibility.

Improved Evaluation Metrics: Developing better evaluation metrics that align with human judgment and preferences. Capturing the quality, relevance, and creativity of captions more accurately.

Real-World Applications: Focusing on real-world applications such as image description for the visually impaired, visual question answering, and image retrieval tasks to make image captioning more practical and useful.

Handling Rare and Unseen Concepts: Addressing the challenge of generating captions for images containing rare or unseen objects or scenes, where the model lacks sufficient training data.

Exploring future directions will advance the state-of-the-art in image captioning, making it more accurate, diverse, and applicable to various real-world scenarios.

Conclusion

The literature survey on image captioning reveals a dynamic and extensive landscape of research in this field. Researchers from computer vision and natural language processing domains are captivated by the task of generating descriptive and contextually relevant captions for images. The survey showcases the evolution of image captioning techniques, from early CNN-RNN models to advanced architectures with attention mechanisms and multimodal fusion. Reinforcement learning, like self-critical sequence training, enhances caption diversity and evaluation metrics. Various methods have been proposed to address challenges like ambiguity and diversity in image captioning. Despite improvements due to deep learning models, evaluating their performance remains complex. Existing evaluation metrics have limitations, leading to ongoing research for more robust alternatives. Looking forward, promising directions for future research also gain significance in image captioning models.

Conflict of Interest

Authors do not have any conflict of interest academic or financial for publication of this work.

References

1. C.P. Chaudhari, S. Devane. Captioning the images: A deep analysis. In *Advances in Intelligent Systems and Computing*; Iyer, B., Nalbalwar, S., Pathak, N., Eds.; Springer, Singapore, **2018**; Vol. 810, pp 987–1000.
2. K. Xu, J.L. Ba, R. Kiros, et al. Show, attend and tell: Neural image caption generation with visual attention. In *32nd International Conference on Machine Learning, ICML 2015*; **2015**; Vol. 3, pp 2048–2057.
3. J. Lu, C. Xiong, D. Parikh, R. Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*; **2017**; Vol. 2017-January, pp 3242–3250.
4. P. Anderson, X. He, C. Buehler, et al. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*; **2018**; pp 6077–6086.

5. S.J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, V. Goel. Self-critical sequence training for image captioning. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*; **2017**; Vol. 2017-January, pp 1179–1195.
6. L. Huang, W. Wang, J. Chen, X.Y. Wei. Attention on attention for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*; **2019**; Vol. 2019-October, pp 4633–4642.
7. J. Cho, J. Lei, H. Tan, M. Bansal. Unifying Vision-and-Language Tasks via Text Generation. In *Proceedings of Machine Learning Research*; **2021**; Vol. 139, pp 1931–1942.
8. J. Gu, J. Cai, S. Joty, L. Niu, G. Wang. Look, Imagine and Match: Improving Textual-Visual Cross-Modal Retrieval with Generative Models; CVPR, **2018**.
9. K. Shuster, S. Humeau, H. Hu, A. Bordes, J. Weston. Engaging image captioning via personality. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*; Manhattan, New York, U.S, **2019**; Vol. 2019-June, pp 12508–12518.
10. L. Guo, J. Liu, P. Yao, J. Li, H. Lu. MSCAP: Multi-style image captioning with unpaired stylized text. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*; Manhattan, New York, U.S, **2019**; Vol. 2019-June, pp 4199–4208.
11. L. Zhou, H. Palangi, L. Zhang, et al. Unified vision-language pre-training for image captioning and VQA. In *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*; **2020**; pp 13041–13049.
12. Y.C. Chen, L. Li, L. Yu, et al. UNITER: Universal Image-text Representation Learning. In *Proceedings of the European Conference on Computer Vision (ECCV*; **2020**; pp 387–404.
13. X. Yao, Y. Peng, J. Chen, X. Zhao, J. Gu. ImageBERT: Cross-Modal Pre-training with Large-scale Weak-supervised Image-text Data. In *Proceedings of the European Conference on Computer Vision (ECCV*; **2020**; pp 41–58.
14. X. Lian, J. Wang, J. Li, et al. Object-Semantics Aligned Pre-training for Vision-Language Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; **2021**; pp 11689–11698.
15. G. Luo, L. Cheng, C. Jing, C. Zhao, G. Song. A thorough review of models, evaluation metrics, and datasets on image captioning. *IET Image Process.* **2022**, 16 (2), 311–332.
16. J. Zhang, K.J. Shih, A. Elgammal, A. Tao, B. Catanzaro. Graphical contrastive losses for scene graph parsing. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*; **2019**; Vol. 2019-June, pp 11527–11535.
17. C. Sun, A. Myers, C. Vondrick, K. Murphy, C. Schmid. Meshed-Memory Transformer for Image Captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV*; **2019**; pp 5414–5423.
18. A. Depeursinge, D. Racoceanu, J. Iavindrasana, et al. Fusing Visual and Clinical Information for Lung Tissue Classification in HRCT Data. *Artif. Intell. Med.* **2010**, 143, ARTMED1118.
19. W. Jiang, X. Li, H. Hu, Q. Lu, B. Liu. Multi-Gate Attention Network for Image Captioning. In *IEEE Access*; **2021**; Vol. 9, pp 69700–69709.
20. T. Yao, Y. Pan, Y. Li, T. Mei. Exploring visual relationship for image captioning. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer International Publishing, Manhattan, New York, USA, **2018**; Vol. 11218 LNCS, pp 711–727.
21. T.N. Kipf, M. Welling. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*; **2017**.
22. S. Hochreiter, J. Schmidhuber. Long Short-Term Memory. *Neural Comput.* **1997**, 9 (8), 1735–1780.
23. S. Ren, K. He, R. Girshick, J. Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*; MIT Press, Cambridge, MA, USA, **2017**; Vol. 39, pp 1137–1149.
24. R. Krishna, Y. Zhu, O. Groth, et al. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *Int. J. Comput. Vis.* **2017**, 123 (1), 32–73.
25. O. Vinyals, A. Toshev, S. Bengio, D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*; **2015**; Vol. 07-12-June-2015, pp 3156–3164.
26. J. Mao, W. Xu, Y. Yang, J. Wang, A.L. Yuille. Neural Image Captioning with Visual Attention. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*; **2015**; pp 129–137.
27. J. Lu, C. Xiong, D. Parikh, R. Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*; **2017**; Vol. 2017-January, pp 3242–3250.
28. O. Press, A. Bar, B. Bogin, J. Berant, L. Wolf. Language Generation with Recurrent Generative Adversarial Networks without Pre-training. In *Proceedings of the International Conference on Learning Representations (ICLR*; **2017**.
29. Y. Feng, L. Ma, W. Liu, J. Luo. Unsupervised image captioning. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*; **2019**; Vol. 2019-June, pp 4120–4129.
30. T.Y. Lin, M. Maire, S. Belongie, et al. Microsoft COCO: Common objects in context. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer, Springer International Publishing, Manhattan, New York, USA, **2014**; Vol. 8693 LNCS, pp 740–755.
31. P. Young, A. Lai, M. Hodosh, J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguist.* **2014**, 2 (67–78), 67–78.
32. P. Sharma, N. Ding, S. Goodman, R. Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*; Long Papers; Association for Computational Linguistics, Stroudsburg, PA, USA, **2018**; Vol. 1, pp 2556–2565.
33. C. Gan, Z. Gan, X. He, J. Gao, L. Deng. StyleNet: Generating attractive visual captions with styles. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*; IEEE Computer Society, Los Alamitos, CA, USA, **2017**; Vol. 2017-January, pp 955–964.
34. K. Papineni, S. Roukos, T. Ward, W.J. Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*; **2002**; pp 311–318.
35. A. Lavie, A. Agarwal. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*; **2007**; pp 228–231.
36. R. Vedantam, C.L. Zitnick, D. Parikh. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*; **2015**; Vol. 07-12-June-2015, pp 4566–4575.
37. C.Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the workshop on text summarization branches out (WAS 2004)*; **2004**; pp 25–26.
38. P. Anderson, B. Fernando, M. Johnson, S. Gould. SPICE: Semantic Propositional Image Caption Evaluation. In *European Conference on Computer Vision (ECCV*; **2016**; pp 382–398.
39. T. Ghandi, H. Pourreza, H. Mahyar. Deep Learning Approaches on Image Captioning: A Review. *A Rev. Vis. Pattern Recognit.* **2022**.