Computer Modeling in Engineering & Sciences

Tech Science Press

Doi:10.32604/cmes.2025.059192

REVIEW

A Survey on Enhancing Image Captioning with Advanced Strategies and Techniques

Alaa Thobhani[1,*], Beiji Zou[1], Xiaoyan Kui[1,*], Amr Abdussalam[2], Muhammad Asim[3], Sajid Shah[3] and Mohammed ELAffendi[3]

[1]School of Computer Science and Engineering, Central South University, Changsha, 410083, China
[2]Electronic Engineering and Information Science Department, University of Science and Technology of China, Hefei, 230026, China
[3]EIAS Data Science Lab, College of Computer and Information Sciences, Prince Sultan University, Riyadh, 11586, Saudi Arabia

*Corresponding Authors: Alaa Thobhani. Email: althobhanialaa@gmail.com; Xiaoyan Kui. Email: xykui@csu.edu.cn

Received: 30 September 2024; Accepted: 31 December 2024; Published: 03 March 2025

ABSTRACT: Image captioning has seen significant research efforts over the last decade. The goal is to generate meaningful semantic sentences that describe visual content depicted in photographs and are syntactically accurate. Many real-world applications rely on image captioning, such as helping people with visual impairments to see their surroundings. To formulate a coherent and relevant textual description, computer vision techniques are utilized to comprehend the visual content within an image, followed by natural language processing methods. Numerous approaches and models have been developed to deal with this multifaceted problem. Several models prove to be state-of-the-art solutions in this field. This work offers an exclusive perspective emphasizing the most critical strategies and techniques for enhancing image caption generation. Rather than reviewing all previous image captioning work, we analyze various techniques that significantly improve image caption generation and achieve significant performance improvements, including encompassing image captioning with visual attention methods, exploring semantic information types in captions, and employing multi-caption generation techniques. Further, advancements such as neural architecture search, few-shot learning, multi-phase learning, and cross-modal embedding within image caption networks are examined for their transformative effects. The comprehensive quantitative analysis conducted in this study identifies cutting-edge methodologies and sheds light on their profound impact, driving forward the forefront of image captioning technology.

KEYWORDS: Image captioning; semantic attention; multi-caption; natural language processing; visual attention methods

1 Introduction

Image captioning serves the purpose of conveying an image's essence in a concise and accurate sentence. It is very similar to employing a machine for translation. Most image captioning models predominantly rely on deep learning [1–3] and neural networks architectures. However, in image captioning, the machine's task is to translate an image into text [4–6]. This involves a profound visual understanding of the image before crafting a meaningful description, one that not only identifies the objects and their attributes within the image but also illustrates the relationship between these objects and the context, be it a location or an activity, the general image captioning model is illustrated in Fig. 1.

Copyright © 2025 The Authors. Published by Tech Science Press.
This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

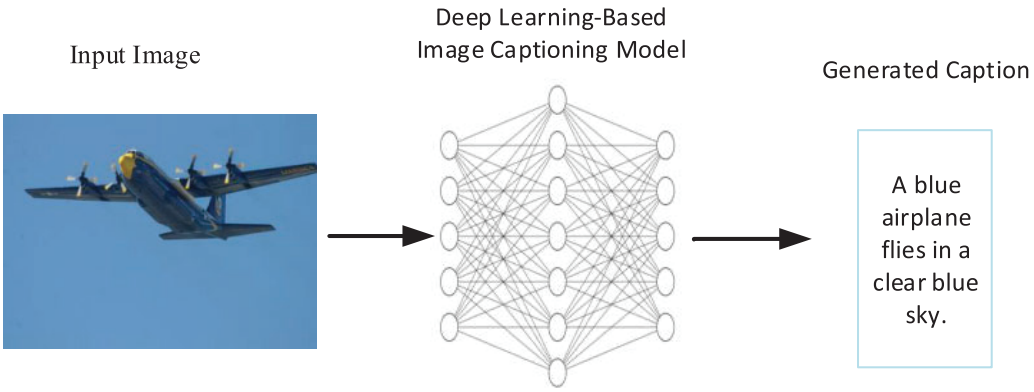

Figure 1: An example of an image captioning model based on deep learning, which processes an input image and produces a corresponding textual description

Image captioning has a wide range of applications, acting as a digital companion for people with visual impairments by offering succinct image descriptions during internet browsing. It enhances autonomous vehicle navigation by providing contextual information for safe, rapid, and precise movement and supports medical diagnosis by generating captions for diagnostic images, aiding timely treatment [7,8]. In news media, automatic captions enrich articles by making visual content more informative [9–11], while in social media, captions boost accessibility and engagement by describing visual posts. In e-commerce, image captioning improves product searchability and user experience with automatically generated descriptions. It also aids search engines by enabling accurate text-based retrieval of visual content. Captioning further extends to service robotics [12–14], military applications [15–17], education [18], geology [19], construction scenes [20], and image indexing [21–23], demonstrating the versatility and value of precise, contextually rich image captions. These varied applications underscore the importance of ongoing research to further enhance real-world implementation.

Many surveys on this topic have been published, offering comprehensive evaluations of image captioning approaches. These include [24–28], which discuss various early methods, as well as more recent works such as [29–33], which focus on advances in neural networks and diffusion models. Additionally, Refs. [34–38] provide in-depth reviews of the latest trends and innovations in the field. In contrast, our work provides a distinctive viewpoint by concentrating on improving the generation of image captions through the identification of the most crucial strategies and techniques within this domain. This survey provides a comprehensive exploration of image captioning strategies and techniques, aiming to enhance the generation of image captions. It covers a variety of aspects, including methods for visual attention, types of semantic information in image captioning, multi-caption generation techniques, neural architecture searches for images, few-shot image captioning, multi-phase learning, and the integration of cross-modal embedding within image caption networks. Additionally, it addresses learning approaches, datasets, and evaluation metrics used in image captioning. By providing a quantitative assessment of state-of-the-art methodologies, this work identifies significant technological advancements, contributing to ongoing progress in the field. It also offers valuable insights to researchers and practitioners by assessing critical methodologies and their impact on improving image captions' quality. Recognizing the evolving technological landscape and its potential for further innovations underscores the paper's relevance for future research and transformative applications in image captioning, cementing its significance.

This paper is organized into key sections covering foundational image captioning frameworks, enhancement strategies, learning approaches, datasets, and evaluation metrics. Studies are analyzed based on

their impact on improving caption generation, innovation, and contribution to advancing state-of-the-art techniques. The evaluation framework focuses on the effectiveness of various methods like visual attention, semantic integration, multi-caption generation, and neural architecture search. Significant advancements are highlighted by systematically reviewing the most influential and cutting-edge research in the field.

2 Image Captioning Frameworks

The evolution of image captioning techniques has seen a shift from traditional encoder-decoder frameworks to modern transformer-based frameworks, each with distinct advantages. Traditional techniques rely on convolutional neural networks (CNNs) for image encoding and recurrent neural networks (RNNs) for text decoding, translating visual features into language in a linear manner. While effective, these models are often limited in handling complex visual scenes and capturing long-range dependencies within captions. In contrast, modern transformer-based frameworks introduce attention mechanisms that interpret relationships among image elements, treating them as contextually interrelated rather than isolated. By leveraging self-attention, transformers generate contextually rich, coherent captions that better capture nuanced scenes, even with complex or lengthy descriptions. These transformer-based models adapt more effectively to diverse contexts and tasks, yielding superior accuracy and detailed captions compared to traditional approaches.

2.1 Traditional Encoder-Decode-Based Frameworks

Presently, the research focus has gravitated toward the automatic description of visual content in images using natural language. Image captioning is an intricate task in the domain of deep learning [39], distinct from other fields like object detection [40–43], recognition [44,45], and image and video tagging [46–50], as it necessitates object detection and the linguistic expression of their relationships.

Traditional encoder-decoder frameworks play a central role in developing image captioning models, notably advancing this field, as emphasized by Vinyals et al. [51]. In this structure, CNN-based encoders convert images into features, working with an RNN-based decoder that transforms these features into descriptive captions, Fig. 2. Typically, the encoder extracts image features to capture the overall visual content, while the captions elaborate on various aspects of the image.

The UpDown [52] framework holds a significant position in image captioning and is often used as a baseline model for numerous image captioning techniques. It generally consists of two Long Short-Term Memory (LSTM) layers: the language LSTM and the attention LSTM. Additionally, it incorporates a visual attention module that focuses on the local features of images as it predicts the next word. This visual attention module selects and emphasizes the crucial regions of an image for generating the subsequent word, as illustrated in Fig. 3.

In previous work, the Deep Hierarchical Encoder-Decoder Network (DHEDN) was introduced, comprising three LSTM layers: the Sentence-LSTM (S-LSTM), Vision-Sentence Embedding LSTM (VSE-LSTM), and Semantic Fusion LSTM (SF-LSTM). The S-LSTM acts as an encoder for the input caption, the VSE-LSTM merges and maps encoded caption features and visual features from the CNN into a shared semantic space, and the SF-LSTM serves as a decoder, generating image descriptions. The distribution combine module and semantic enhancement module were incorporated into the SF-LSTM to enhance its capacity. The policy gradient method was applied to optimize this model [53].

In another previous study, the Recall Network framework was introduced as an encoder-decoder-based model that consistently retrieves the image's visual information during the generation of each word. This framework, implemented using the GridLSTM, adjusts to the image's visual representations through

the depth of LSTM's memory cells. It efficiently updates and forgets visual information in line with the corresponding word without requiring additional learnable parameters or networks [54].

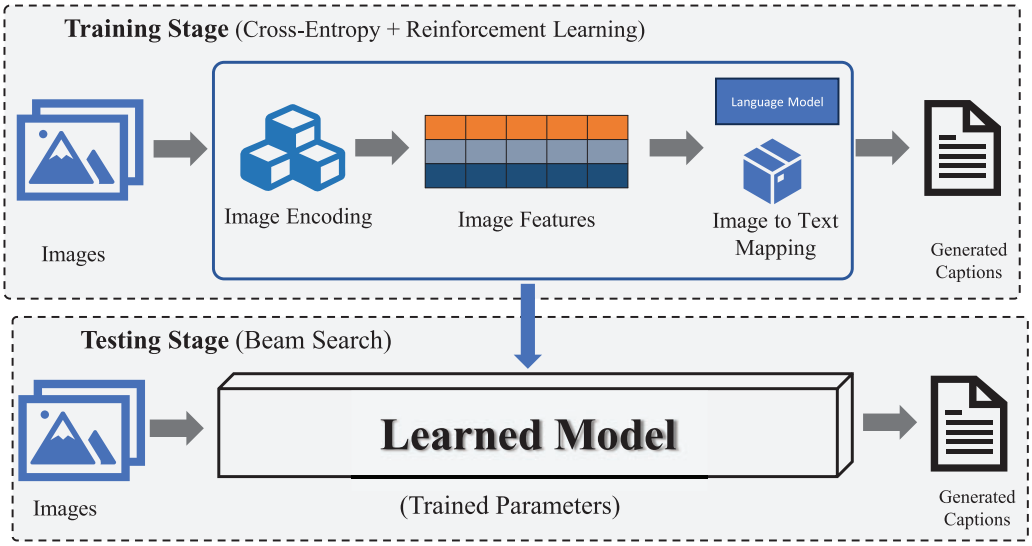

Figure 2: Illustration of the traditional encoder-decoder framework

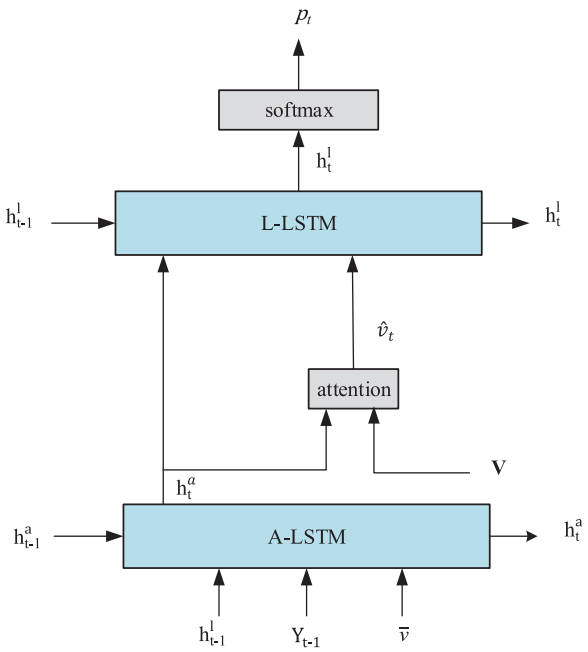

Figure 3: A two-layer LSTM with attention, inspired by the bottom-up, top-down attention mechanism. The A-LSTM represents the attention LSTM, responsible for modeling attention over the extracted features (V), while the L-LSTM refers to the language LSTM, which generates the output sequence. The term $\bar{v}$ refers to the global features

2.2 Transformer-Based Frameworks

Transformer-based frameworks, as illustrated in Fig. 4, have achieved significant advancements in image captioning, building on their success in natural language processing and extending it to computer vision. These deep learning models leverage a powerful self-attention mechanism to capture intricate relationships across input sequences, whether in text or visual data, without relying on recurrence as in traditional models (like RNNs). This self-attention enables transformers to attend to all parts of an input simultaneously, making them highly efficient for large datasets. In image captioning, transformer architectures, including Vision Transformers, are particularly effective at modeling long-range dependencies, allowing for detailed and contextually accurate descriptions of complex visual scenes and delivering state-of-the-art performance across various AI applications. Transformer-based frameworks improve image captioning models' performance by encoding spatial and contextual information more effectively, resulting in more accurate and contextually rich captions. In their seminal work, Vaswani et al. [55] introduced the fully-attentive paradigm, which brought a revolutionary perspective to language generation. This innovative approach, using the Transformer model as its foundation, has significantly impacted various natural language processing tasks, including GPT (Generative Pre-training Transformer) [56] and BERT (Bidirectional Encoder Representations from Transformers) [57], playing pivotal roles in reshaping the field. The Transformer design shares similarities with image captioning, as it adopts a sequence-to-sequence structure, making it a suitable solution for such tasks. In the conventional Transformer, masking self-attention is employed to process words, serving the role of a decoder. Further, cross-attentional processing is applied, where words function both as queries and as outputs from the final encoder layer, followed by a feed-forward network. During training, the preceding words are masked, facilitating unidirectional generation. Some image captioning models have employed the standard Transformer architecture [58,59].

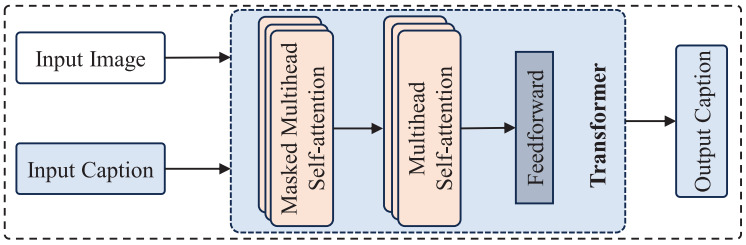

Figure 4: Illustration of the transformer-based frameworks

The Bidirectional Encoder Representations from Transformers (BERT) was primarily designed for Natural Language Processing (NLP). By achieving bidirectional context understanding through the Masked Language Model (MLM) objective, BERT outperforms unidirectional models in various Natural Language Processing (NLP) benchmarks. The concept has become a fundamental part of the development of transformer-based models for natural language understanding and generation. However, its contextualized embeddings and bidirectional attention mechanism have inspired adaptations for multimodal applications, including image captions. For improved image understanding and caption generation, researchers have explored hybrid models that incorporate BERT-like structures in conjunction with vision models. These innovative architectures seamlessly integrate visual and textual modalities at early stages. The primary advantage of this approach, when used to model captioning, lies in the initialization of text-related layers that have been trained using parameters that are derived based on extensive corpora of text. Consequently, BERT has found substantial application in works utilizing pre-training techniques [60,61]. For instance,

Ref. [61] developed a captioning algorithm that leverages both the textual and visual aspects within a BERT-based model [62]. Moreover, in [60], the utilization of image object markers for anchoring was introduced to enhance alignment of language and vision collective representations. To achieve this objective, the model encodes the image and text pair as a triple, composed of tokens that represent words, tags that represent objects, and features of regions. The object attributes are derived from the object detectors of the textual classes, providing a rich foundation for connecting language and visual context in captioning tasks.

The pre-training of vision-language models often involves the utilization of transformer-based architectures. Transformer-based models have emerged as the cornerstone of many state-of-the-art vision-language models due to their success in a variety of natural language processing tasks. Ref. [63] presents the Vision-Language Pre-training (VLP) model, a versatile encoder-decoder framework tailored for vision-language tasks. VLP employs a mutually shared transformer network, initially trained on a large caption dataset. It enhances performance through masking bidirectional and sequence-to-sequence language predictability tasks, accommodating context variations, and improving its applicability to tasks such as image captioning and visual question answering. Ref. [64] introduces mPLUG, a unified Multi-modal Pre-training framework for both vision-Language Understanding and Generation, a vision-language model that tackles challenges like computational inefficiency and information disparity using innovative cross-modal skip-connections. The connections establish shortcuts between layers, enhancing efficiency, particularly in visual self-attention mechanisms. Additionally, Ref. [65] introduces Prismer, a vision-language model that leverages domain expertise and excels in fine-tuned and few-shot learning scenarios, requiring minimal training data. Ref. [66] presents the Qwen-VL series, a set of multilingual models of visual language that excel in multimodal understanding and generation.

GPT models, primarily designed for natural language processing, can be repurposed for image captioning through a multimodal approach. By combining GPT with a vision encoder to extract visual features from images, the model can generate captions based on textual prompts and visual information. This adaptation typically involves fine-tuning the pre-trained GPT model on image-caption pairs to optimize its performance for this task. While not initially intended for image captioning, the flexibility of GPT enables it to handle multimodal tasks effectively, although specialized models may outperform it due to their tailored architectures. Ref. [67] introduces VisualGPT, which leverages linguistic knowledge from a pre-trained language model (LM) to balance visual and linguistic information, enabling quick adaptation to new tasks with minimal in-domain training data. The model employs a unique self-resurrecting encoder-decoder attention mechanism that efficiently integrates the pre-trained LM as the language decoder, yielding sparse activations without encountering zero gradients. XGPT [68] presents a new method for Cross-modal Generative Pre-Training in image captioning. XGPT employs four innovative generation tasks—Adversarial Image Captioning (AIC), Image-conditioned Masked Language Modeling (IMLM), Image-conditioned Denoising Autoencoding (IDA), and Text-conditioned Image Feature Generation (TIFG)—to pre-train text-to-image caption generators.

The work presented by [69] introduces a method by leveraging the Contrastive Language-Image (CLIP) model for visual encoding and a pre-trained language model (GPT2) for caption generation. By fine-tuning a language model on CLIP-encoded captions, the method efficiently produces meaningful captions for diverse datasets without additional annotations or pre-training. Ref. [70] resides in Visual Conditioned GPT (VC-GPT), a comprehensive image captioning framework that streamlines the conventional two-stage training paradigm. VC-GPT seamlessly links a pre-trained visual encoder (CLIP-ViT) with a language decoder (GPT2), thereby eliminating the necessity for distinct object detection models. Ref. [71] presents PromptCap, a knowledge-based visual question answering (VQA). PromptCap utilizes a natural-language prompt to govern the visual elements depicted in the resulting caption, thereby improving the interaction between

images and large language models such as GPT-3. Ref. [72] introduces SmallCap that addresses challenges associated with scaling data and model size. SmallCap produces captions using an input image and associated captions obtained from a datastore. It is characterized by its lightweight nature and quick training process, as it only updates parameters in cross-attention layers connecting a pre-trained CLIP encoder with a GPT-2 decoder. By substituting the contents of the data store, SmallCap adapts to new domains seamlessly without needing further fine-tuning.

3 Strategies and Techniques

This section explores key strategies and techniques that significantly enhance image caption quality and context. This dynamic field requires innovative approaches, including attention mechanisms, semantic-rich captions, multi-caption generation, neural architecture search, few-shot, and cross-modal embedding. These strategies elevate caption coherence, informativeness, and creative depth, ultimately empowering image captioning systems for various applications. The taxonomy of strategies to improve image captioning generation is shown in Fig. 5, and Table 1.

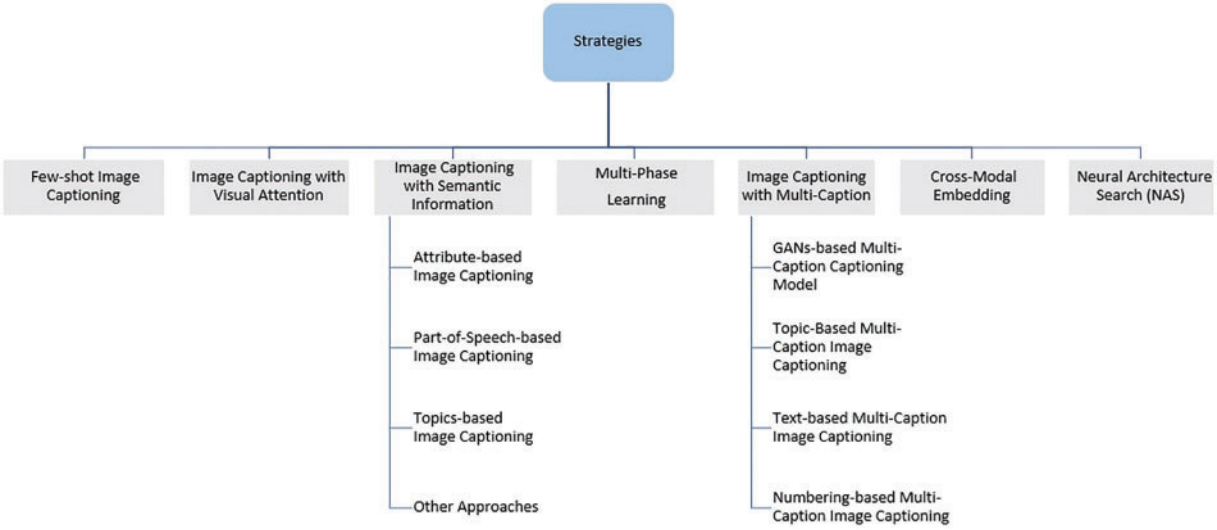

Figure 5: Taxonomy of strategies to improve image captioning generation

Table 1: A list of some strategies and techniques adopted in some papers in this survey

Model	Ref.	Technique used	Category of the strategy used
Hard + Soft attention	2015, [73]	Image captioning with visual	Visual attention
ATT-FCN	2016, [58]		
UpDown	2018, [52]		
X-Linear attention	2020, [75]		
hLSTM + Adaptive Attention	2020, [76]		
BCAN	2021, [74]		
JRAN	2023, [96]		
AbSViT	2023, [86]		
GVA	2024, [88]		
CGAN	2017, [109]	GANs-based	Multi-caption
TOMS	2018, [110]	Topics-based	
TextCap	2021, [111]	Text-based image captioning	

(Continued)

Table 1 (continued)

Model	Ref.	Technique used	Category of the strategy used
NumCap	2023, [112]	Numbering-based image captioning	Semantic information
NeuralTalk2-T-oe	2018, [101]	Topics-based image captioning	
TopicSensitive	2020, [104]		
TopicBased	2020, [105]		
ETP + RTR + STP	2023, [106]		
MAD + SAP	2020, [90]	Attribute-based image captioning	
Stack-VS	2020, [93]		
Fusecap	2023, [94]		
FaceAtt	2023, [95]		
PoS + Guidance	2020, [99]	Part-of-Speech-based Image Captioning	
PoSGuidance + multimodal	2022, [98]		
NPoSC-A3	2024, [100]		
Re-Caption	2019, [107]	Attribute-based with Part-of-Speech	
AutoCaption (Vinvl)	2020, [113]	Neural architecture search image captioning	Neural Architecture Search (NAS)
MMnas	2020, [114]		
IUC	2022, [115]		
Natural-language-feedback	2017, [116]	Multi-phase learning image captioning	Multi-phase learning
Stack-captioning	2018, [117]		
Re-Caption	2019, [107]		
FPAIT	2018, [118]	Few-shot image captioning	Few-shot
LMCap	2023, [119]		
Re-ViLM	2023, [120]		
VSE++	2017, [121]	Visual-semantic embedding	Cross-modal embedding
NeuralTalk2-T-oe	2018, [101]		
MLBL-F	2014, [122]		
ORDER	2015, [123]		
MAP + NDCG	2019, [124]		
Joint corr-learning	2019, [125]		
Oscar	2020, [126]		
CLIP	2021, [127]		

3.1 Image Captioning with Visual Attention

Attention mechanisms have been introduced to enhance image captioning by effectively incorporating image features [58,73,74]. These mechanisms, in the context of visual attention, employ image features that are dynamically weighted to predict subsequent words during caption generation at each time step. Various mechanisms of attention have been integrated to improve image descriptions [75,76].

Attention mechanisms have been introduced to enhance image captioning by effectively incorporating image features [58,73,74]. These mechanisms, in the context of visual attention, employ image features that are dynamically weighted in order to predict subsequent words during caption generation at each time step. A variety of attention mechanisms have been integrated to improve image descriptions [75,76].

In previous studies, attention mechanisms were integrated into image captioning in order to guide the network's attention to different image regions while predicting words [73]. Some frameworks introduced adaptive attention, allowing the model to determine whether to rely on language model or visual features when generating words, using a sentinel gate [77]. Another approach combined bottom-up and top-down

visual attention mechanisms, incorporating object proposals generated by a Faster-RCNN with a two LSTM layers-based captioning network [52]. The attention mechanism was further enhanced by integrating spatio-temporal memory mechanisms into visual attention, creating the spatio-temporal memory attention model (STMA) [78]. Additionally, a task-adaptive attention module was designed to reduce the model's dependency on visual information when generating non-visual words [79].

To improve caption quality and explore visual-semantic relationships, a dual attention approach to pyramid image feature mapping was applied [80]. A cluster-based grounding network was introduced to focus on linguistic characteristics within informative regions without introducing additional parameters or raising inference complexity [81]. Furthermore, the Proposal Attention Correctness (PAC) metric was introduced to assess the correctness of soft attention, bridging the gap between captioning performance and visual grounding. An extension of the Transformer model, the multimodal Transformer, emphasized both inter-modal and intra-model interactions within a single attention block [82]. It enabled modeling of word-to-object, object-to-object, and word-to-word relationships, along with multi-view feature learning. Finally, region-aware interaction learning was proposed to extract information from object and region dimensions, considering semantic correlations between them [83]. Ref. [84] introduces an innovative image caption generation approach, incorporating wavelet decomposition and convolutional neural networks for comprehensive information extraction. Ref. [85] presents a new method for reevaluating attention mechanisms in vision transformers, challenging the traditional interpretation and introducing the concept of similarity grouping. Ref. [86] introduces an innovative approach to top-down attention, drawing inspiration from the Analysis-by-Synthesis (AbS) theory of human vision. It introduces the Analysis-by-Synthesis Vision Transformer (AbSViT) model, which not only approximates AbS but also offers controllable top-down attention. The work presented in [87] introduces a Refined Visual Attention (RVA) framework for image captioning that dynamically reweights visual attention by considering the context provided by previously generated words. Similarly, Ref. [88] unveils a Guided Visual Attention (GVA) methodology for generating image captions, which fine-tunes attentional weights to enhance the overall quality of the captions produced. These methodologies represent pivotal progress in leveraging attention mechanisms to create more accurate and contextually relevant captions.

Despite their innovative solutions, these models still face significant challenges. A major issue is their tendency to overly focus on local features, often at the expense of a comprehensive understanding of the image. This can lead to captions that, while technically accurate, fail to capture the main idea or highlight the most important aspects of the image. Additionally, these models struggle with varying visual presentations of objects, such as changes in scale, position, and occlusions. These variations can significantly hinder the models' ability to accurately identify and interpret the key elements of an image, thereby affecting the relevance and accuracy of the generated captions.

3.2 Image Captioning with Semantic Information

Incorporating image attributes, which refer to the most common words in the training caption corpus, has been explored as an approach to enhance image captioning, leveraging the rich semantic information carried by these attributes [89–92]. Prior studies integrated attribute detectors with images to predict these attributes and assigned relative probability distributions. Attributes predicted from image features were then combined together with image features and used as inputs for caption generation, enabling caption generation to incorporate latent semantic image information, thereby improving caption quality.

3.2.1 Attribute-Based Image Captioning

One study introduced a multimodal attribute detector that effectively utilizes image attributes [90]. This detector, in conjunction with the Faster-RCNN module, obtained visual object features from images and predicted their attributes, Fig. 6 illustrates an example of the local region features extracted using the Faster-RCNN. By training the multimodal attribute detector alongside the captioning model, both modules cooperatively leveraged attribute semantics. Moreover, this approach involved selecting a group of attributes closely related to the current linguistic context at each time step for generating the forthcoming caption word. A multi-stage image descriptor, known as Stack-VS, was introduced to efficiently utilize semantic and visual-level information from input images [93]. This model employed bottom-up and top-down attention techniques within a stack decoder model, optimizing attention weights for semantic-level attribute representations and visual-level feature representations to produce detailed image descriptions. Visual-level features were obtained using a Faster-RCNN network, and a weakly-supervised multiple-instance learning approach was employed to learn the semantic-level attributes of input images. Ref. [94] presences FUSECAP, enriching captions with visual expert insights and a large language model, creating 12 million improved caption pairs. These enhanced captions improve image captioning models and benefit image-text retrieval. Ref. [95] introduces the Face-Att model, focusing on generating attribute-centric image captions with a special emphasis on facial features. JRAN [96] enhances caption coherence by focusing on feature relationships. By incorporating both region and semantic features, JRAN generates precise and context-aware captions, offering valuable contributions to applications like chatbots and image search.

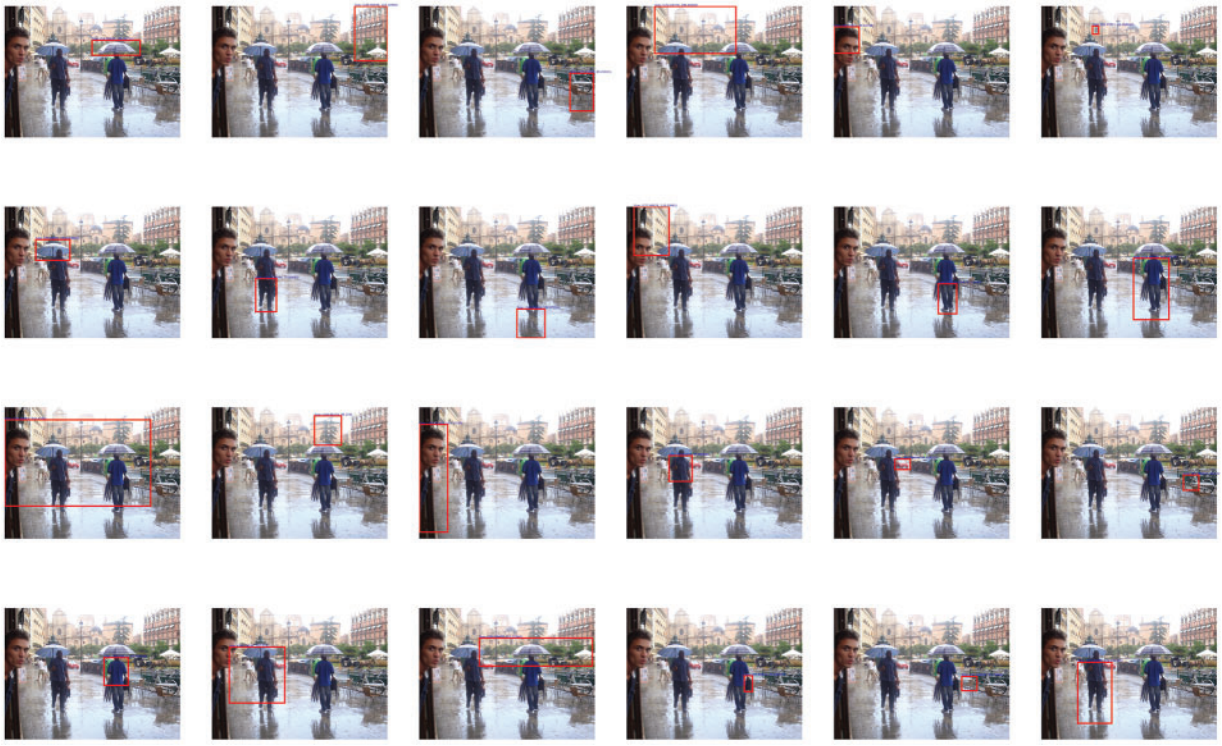

Figure 6: (Continued)

Figure 6: An example image showing the top 36 local region features extracted using Faster-RCNN [97]

3.2.2 Part-of-Speech-Based Image Captioning

Several recent studies incorporated part-of-speech (PoS) information into image captioning models, introducing methods to control the flow of information based on predicted PoS [98]. These approaches, such as the PoS guidance module, combined PoS information with word embeddings to influence caption generation [98]. Additionally, some models incorporated PoS into the encoder-decoder framework, allowing it to direct the sentence generation task by detecting and examining PoS tags for upcoming words [99]. The PoS predictor was fused with image descriptors using different styles, namely the inject-based and merge-based styles, to collaboratively guide the text generation process, with multi-task learning introduced for effective training. NPoSC-A3 [100] mechanism stands out by effectively leveraging PoS clues to judiciously integrate both visual and semantic information into the language model. With its unique components, including a global semantic context generator and a PoS predictor, NPoSC-A3 demonstrates significant improvements in the accuracy and syntactic richness of the generated captions.

3.2.3 Topics-Based Image Captioning

Other studies explored the integration of topics with image captioning algorithms, utilizing topic models with caption corpora. For instance, one approach incorporated topics to direct sentence generation by detecting the top-1 topic for an input image and embedding it alongside the image [101]. Another introduced a topic-guided image descriptor that detected relevant topics based on image content, allowing topic-based re-weighting to enhance image descriptions. Ref. [102] presences an attention mechanism called topic-guided attention for image captioning. The proposed mechanism integrates image topics as guiding information, enhancing the selection process. Additionally, the method employs separate networks to extract image features and topics, allowing for joint fine-tuning in an end-to-end manner during training. Ref. [103] introduces a new topic-guided neural image captioning model, which improves upon existing methods that often overlook high-level semantic information. Zia et al. [104] proposed framework generates captions that are sensitive to the topics depicted in the images, capturing the semantic relations and polysemous nature of words. Ref. [105] emphasizes on incorporating the topic or purpose of an image into the caption generation process beyond just object detection and orientation. By focusing on understanding the broader context of images, the proposed deep learning framework aims to generate captions that better reflect human-like perception. This results in superior descriptions that are not only grammatically correct but also contextually relevant to the content of the images. The proposed model incorporates a topic model into the framework, representing images as sets of topics with relevant word distributions. In the study by Al-Qatf et al. [106], an innovative end-to-end training strategy is introduced, aiming to enhance the alignment

of image captions with image content by integrating topic prediction alongside caption generation. This approach incorporates several key components, including an Enhanced Topic Predictor (ETP), Retrieval-Based Topics Re-weighting (RTR), and a Subsequent Topic Predictor (STP). Through these mechanisms, the model achieves the generation of image captions that are not only more accurate but also notably diverse, contributing to a richer and more comprehensive description of the images. This approach represents a significant stride in the quest for more effective and contextually relevant image captioning techniques.

3.2.4 Other Approaches

While various approaches have been explored, such as combining attributes and part-of-speech (PoS) analysis or integrating topics with PoS information, these efforts have paved the way for significant advancements in image description generation. One notable development is the proposal of a two-phase learning re-captioning model, which aims to further enhance the quality of generated descriptions. This model achieves its goals by capitalizing on saliency mechanisms, including visual, semantic, and sample saliency, to refine image representations [107]. It accomplishes this through the integration of a multi-label Convolutional Neural Network (CNN), bidirectional Gated Recurrent Unit (GRU), attention networks, and a joint embedding layer, resulting in improved image descriptions.

Despite their efficiency, semantic attention techniques in image captioning face several limitations that affect their performance. These limitations result in an inadequate representation of important image features, potentially leading to vague captions or the omission of significant visual details. Additionally, there is a reliance on predefined attribute sets, which limits the system's ability to adapt to new or unexpected visual elements that fall outside these categories. These techniques are also prone to errors in part-of-speech (PoS) tagging. Since semantic attention often relies on PoS information to identify and focus on key elements in an image, incorrect PoS tagging can misdirect the attention mechanism, resulting in less relevant or accurate captions. Another challenge is capturing a wide range of topics without introducing biases from the underlying saliency-detecting methods. Saliency-based techniques often prioritize visually dominant elements, which may not align with the narrative or contextual importance of an image, leading to captions that overlook subtle but critical details. Furthermore, current semantic attention frameworks struggle to dynamically adapt to the varying complexity of visual scenes. As images become more complex, featuring a diverse array of objects, activities, and interactions, it becomes increasingly difficult to accurately capture and translate this complexity into comprehensive and coherent captions [108].

3.3 Image Captioning with Multi-Caption

Crafting a single, exhaustive caption that encapsulates the complexities and subtleties of a complex image is an inherently demanding endeavor, even for human observers. To tackle this challenge, certain image captioning techniques adopt a more inclusive strategy, generating multiple descriptions for a single image. This approach seeks to offer a more thorough and comprehensive portrayal of the image's content, taking into consideration various visual elements, particulars, and potential interpretations. These multiple descriptions can provide a more diverse and in-depth viewpoint on the same image, augmenting its accessibility and comprehension.

3.3.1 GANs-Based Multi-Caption

The GANs-based multi-captioning model takes an innovative approach to describing images by generating several diverse captions [109], each highlighting different features of the image. It does this through two main components: the generator and the evaluator.

The Conditional Generative Adversarial Network (CGAN) for image captioning operates by combining a Generator (G) and an Evaluator (E) in an adversarial framework. The Generator takes image features (extracted from a Convolutional Neural Network, CNN) and a random noise vector (z) as inputs to produce captions word-by-word using a Long Short-Term Memory (LSTM) network. The random noise vector introduces diversity in the generated captions. The Evaluator scores the captions based on their semantic relevance to the image and their naturalness compared to human-written captions. Training is achieved using reinforcement learning with Policy Gradient, where the Evaluator's score serves as a reward signal for the Generator. Additionally, Monte Carlo rollouts provide early feedback during caption generation, helping stabilize and improve the training process. This approach ensures that the generated captions are diverse, natural, and semantically aligned with the images.

3.3.2 Topic-Based Multi-Caption

Another approach leverages the concept of topics by generating multiple captions for a given image [110]. With topic-based multi-caption model, the input consists of an image and a specified topic. Modeling is designed to create captions that are specifically related to the chosen topic. A fusion gate unit is introduced to maintain thematic consistency and effectively incorporate the topic information into the captioning process. This unit enables the model to seamlessly integrate topic embeddings into the Long Short-Term Memory (LSTM) network. Furthermore, a topic classifier is employed to facilitate topic prediction, ensuring that the generated captions remain contextually relevant to the chosen topic [110].

3.3.3 Text-Based Multi-Caption

Another work [111] introduced the Anchor-Captioner approach in text-based image captioning, excelling in fine-grained description tasks by generating multiple captions from various viewpoints. Its novelty lies in its emphasis on fine-grained text information, enabling it to decipher and incorporate textual content from elements like billboards, signs, and product prices within images.

3.3.4 Numbering-Based Multi-Caption

Recently, the model presented in [112] introduced a unique approach that considers the number of available ground truth captions for an image during training. This model learns from these numbers and employs them to generate diverse image captions. Instead of relying solely on the semantic information provided by GT captions, the model takes advantage of the quantitative availability of multiple captions to create a varied set of captions for images, as illustrated in Fig. 7.

These approaches highlight the significance of generating several descriptions of an image, as they enable a more comprehensive and nuanced representation of visual content. This, in turn, can benefit various applications, including image indexing, accessibility, and enhanced user comprehension, by offering a range of perspectives on complex visual data. However, in many contexts, users often prefer a single, well-crafted caption that provides a comprehensive understanding of the image content. While the Multi-Caption approach can be beneficial in certain scenarios, such as when multiple perspectives or interpretations are warranted, the simplicity and clarity of a singular, high-quality caption often better serve the user's need for a clear and concise understanding. A single caption can efficiently convey the essential information without overwhelming the viewer with unnecessary details or redundancy. It allows for a focused presentation of key elements and concepts within the image, facilitating quick comprehension and engagement. Additionally, a well-written single caption can maintain the viewer's attention and interest, guiding them through the visual narrative with precision and effectiveness. Moreover, in scenarios where space or attention is limited, such

as social media posts or mobile applications, a single caption is more practical and user-friendly. It avoids cluttering the interface and ensures that the message is conveyed succinctly and effectively. However, crafting a single caption that strikes the right balance between informativeness and brevity can be a challenge. It requires careful consideration of the most salient aspects of the image and the intended message, as well as skillful writing to convey this information concisely yet comprehensively.

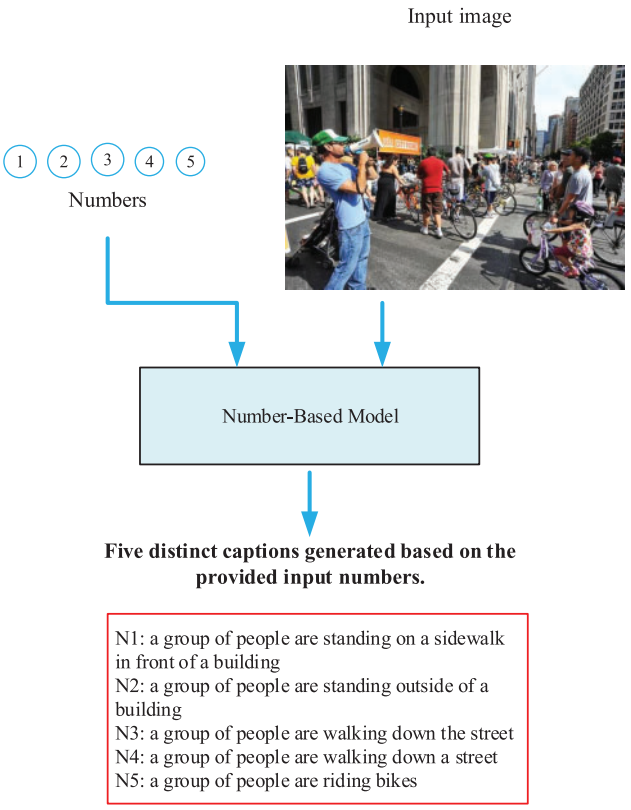

Figure 7: Overview of the numbering-based multi-caption model

3.4 Multi-Phase Learning

Numerous approaches have delved into Multi-Phase Learning in image captioning, reflecting the complexity of the task and the continuous quest for improved performance. For example, Gu et al. [117] proposed a novel coarse-to-fine image captioning model integrating stacked visual attention and multiple LSTM networks. Departing from conventional one-stage models, their approach facilitates gradual caption generation in a coarse-to-fine manner, proving advantageous for image captioning tasks. By iteratively refining the generated captions through multiple phases, the model can capture finer details and nuances in the image content, leading to more accurate and descriptive captions. However, a potential drawback is identified in the sequential arrangement of Gu et al.'s two-phase learning, as the second phase does not commence until the completion of the first phase. This sequential dependency may introduce delays in the caption generation process and limit the model's adaptability to dynamic image features.

In another innovative strategy, Fidler et al. [116] integrate natural language feedback from a human teacher into reinforcement learning for image captioning. By incorporating human feedback into the training process, their approach aims to improve the relevance and coherence of generated captions. They employ a hierarchical phrase-based RNN to accommodate human feedback, enabling the model to learn from explicit

linguistic guidance. While promising, a limitation arises from the costly nature of human language feedback, which slows down the optimization of captioning models and may impose practical constraints on large-scale deployment.

Similarly, in the work by Zhou et al. [107], a saliency-enhanced re-captioning framework is introduced, employing a two-phase learning approach for image captioning. This framework distills visual, semantic, and sample saliency from the first-phase model to enhance global image representation in the second phase. By leveraging saliency cues derived from the initial captioning phase, the model can focus on the most relevant aspects of the image and generate more contextually rich captions. However, a notable aspect is the self-boosting nature within the same model across the two phases, resulting in the generation of only one caption in the output of phase one. While this self-boosting mechanism may lead to more focused attention and enhanced caption quality, it also introduces potential biases and limitations in the diversity of generated captions. Fig. 8 shows a schematic representation of the multi-phase learning framework.

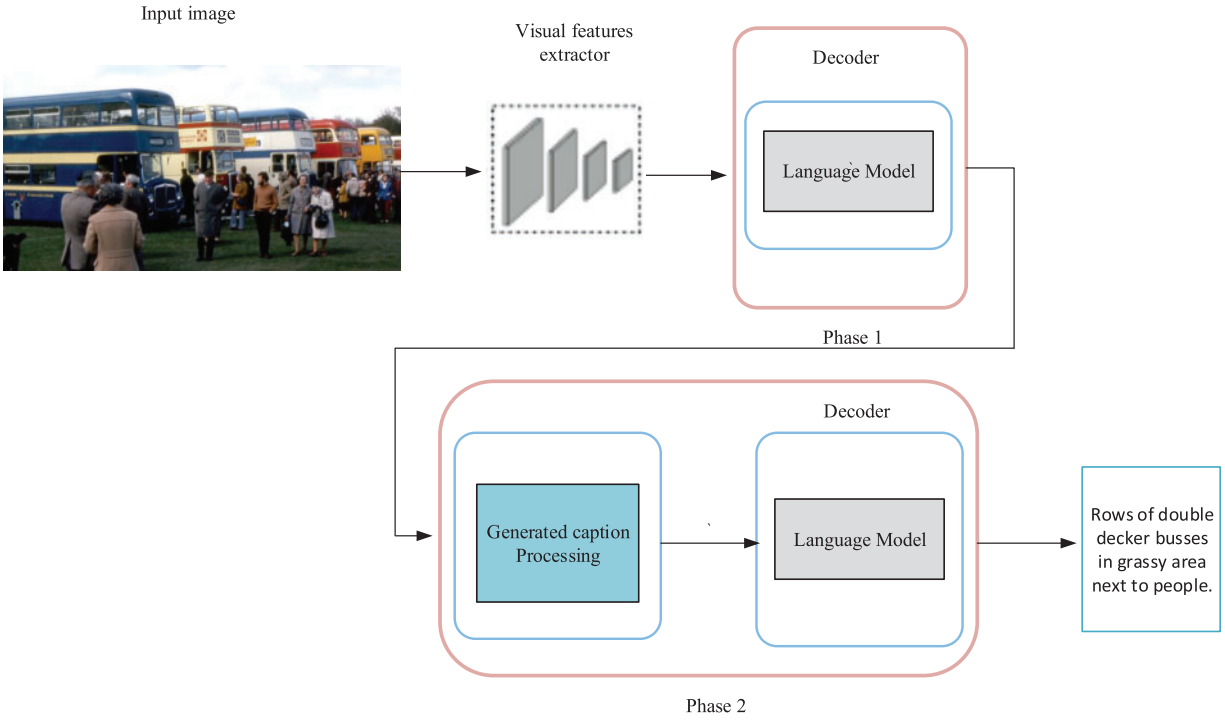

Figure 8: A schematic representation of the multi-phase learning framework, highlighting the two distinct phases

3.5 Neural Architecture Search (NAS)

Neural architecture search (NAS) for image captioning involves automating the process of designing and optimizing neural network architectures specifically tailored for the task of generating captions for images. The goal of this method is to identify architectures that can efficiently represent the intricate connections between text and image, resulting in captions that are more precise and appropriate for the context. The study by [113] introduced AutoCaption, a novel approach to image captioning that utilizes Neural Architecture Search (NAS) to automatically design a more effective decoder for generating image captions. By employing a network structure search method based on reinforcement learning, AutoCaption accelerates the process of designing the text generation model while ensuring it considers the specific characteristics of the image

captioning task. MMnas [114], which stands for Multimodal Neural Architecture Search, is designed to tackle various multimodal learning tasks. MMnas employs a unified backbone consisting of deep encoder-decoder structures. These structures incorporate primitive operations selected from a predefined pool, with task-specific heads added to address different multimodal learning tasks. Through a gradient-based NAS algorithm, MMnas efficiently learn optimal architectures for diverse tasks. A three-tiered strategy for image captioning is proposed by [115]. Using differentiable architecture search techniques, this approach finds the most appropriate architecture automatically. The optimization process unfolds across three sequential stages: first, the image captioning model refines its encoder and decoder weights to generate captions; then, the refined encoder-decoder produces a synthetic captioning dataset from unlabeled images, which allows the model's weights to be updated further; finally, the model evaluates its performance on a validation set and adjusts the encoder-decoder architecture to minimize validation loss.

However, NAS for image captioning faces several challenges. First of all, there can be a large and intricate search space for neural architectures, which makes the exploration process costly and time-consuming in terms of processing. Furthermore, the selection of search strategy, optimization algorithms, and evaluation criteria all have a significant impact on the performance of NAS approaches, necessitating careful design and adjustment. Notwithstanding these difficulties, new architectures that surpass the capabilities of handcrafted designs may be found with NAS, utilizing algorithmic optimization and processing power to push the limits of image captioning performance. NAS can spur innovation in image captioning and make it easier to create more practical and efficient models by automating the architecture design process.

3.6 Cross-Modal Embedding

Cross-modal embedding techniques have garnered substantial attention in a variety of cross-modal studies [122,124,125]. These approaches involve embedding different modalities, such as images and captions, into a shared embedding space. Within this space, the similarity between modalities is measured using metrics like Euclidean or cosine distance, which enable the quantification of their relatedness [122]. Typically, the relationship between modalities is modeled symmetrically, treating them as equals. However, an alternative approach proposed in previous work by [123] explores an asymmetrical treatment of the image-caption relationship. Here, a two-level partial order is introduced, establishing a coordinate-wise order within the common embedding space. This approach harnesses the hierarchical structure of visual-semantic relationships by mapping this partial order to the visual-semantic hierarchy while preserving order. As part of the advancement of cross-modal retrieval, Ref. [121] introduced a novel technique for learning visual-semantic embeddings, enhancing the loss function with the inclusion of hard negatives, resulting in the VSE++ model. Additionally, Ref. [101] extended cross-modal embedding to incorporate topics alongside images and captions, enforcing a three-level hierarchical structure in the embedding space and enabling coordination between topics, captions, and images. Ref. [127] uses contrastive learning to create a shared embedding space for images and text, allowing images to be recognized based on textual descriptions alone. By training on vast datasets of image-text pairs, the model can generalize across tasks without additional fine-tuning, enabling zero-shot learning where it can classify new images based solely on their descriptions. This represents a significant shift from traditional supervised methods that require labeled image datasets. The novelty presented in [126] is the introduction of object tags as anchor points for aligning images and text in a shared semantic space. By using object tags detected in images, the model, named Oscar, simplifies the process of learning semantic alignments between visual and textual elements. This approach enhances pre-training efficiency and improves performance across multiple vision-language tasks, such as image captioning and visual question answering, by providing explicit alignment cues, making it more effective than prior methods that rely solely on self-attention without anchor points.

3.7 Few-Shot Image Captioning

Few-shot image captioning aims to create descriptive texts for images by training a model with a very limited set of examples. This challenge is tackled by leveraging models pre-trained on vast datasets, which are then fine-tuned on the small available dataset to adapt to the specific task of captioning new images. The approach generally involves an image encoder, like a CNN, to extract visual features and a text decoder, such as an LSTM or a Transformer, to generate captions. The key strategies include using transfer learning to apply knowledge from related tasks, employing attention mechanisms to focus on relevant parts of the image, and augmenting the limited data to enhance the model's learning. The process is iterative, with adjustments to the model's architecture and training methods based on evaluation metrics like BLEU and CIDEr, aiming to improve the relevance and accuracy of the generated captions.

The Fast Parameter Adaptation for Image-Text Modeling (FPAIT) [118] introduces a novel method for few-shot learning in multi-modal tasks like image captioning and visual question answering. It stands out by quickly adapting to new tasks with minimal examples and mitigating biases caused by small datasets through dynamic linear transformations. This enables FPAIT to significantly improve performance in understanding the intricate relationships between images and text. Ref. [128] proposes a new approach to few-shot image captioning by utilizing unpaired images and captions, thereby minimizing the need for expensive manual annotation of large-scale datasets. The proposed method, based on ensemble-based self-distillation, involves training multiple base models with different data samples in each iteration. Pseudo captions and features are generated from unpaired data and used to train the base models, with weights allocated based on their confidence levels.

Recent research highlights effective methods of integrating few-shot learning with transformer architectures for image captioning. Research into in-context configurations further explores image selection and caption assignment, offering strategies to improve vision-language model adaptation [129]. LMCap uses a multilingual CLIP encoder to retrieve similar captions, which prompts a language model to create captions in multiple languages without requiring supervised data [130]. Tsimpoukelli et al. [119] present a novel approach aimed at expanding the few-shot learning potential of auto-regressive language models into a multimodal domain that encompasses both visual and linguistic elements. By training a vision encoder to represent images as sequences of embeddings and pairing them with pre-trained language models, the system can generate captions with just a few examples. This approach creates a multimodal few-shot learner capable of rapidly acquiring new tasks when provided with examples represented as sequences of image and text embeddings. The system demonstrates impressive versatility, including learning new words for objects, visual question-answering with minimal examples, and leveraging external knowledge. Ref. [120] introduces Re-ViLM, a Retrieval-augmented Visual Language Model built upon Flamingo, aimed at addressing limitations in existing approaches for image-to-text generation. Re-ViLM retrieves relevant knowledge from an external database for zero and in-context few-shot image-to-text generations. This approach reduces model parameters and allows easy accommodation of new data during evaluation. Moreover, Re-ViLM facilitates in-context few-shot learning by constructing interleaved image and text data. LMCap presents a novel approach to multilingual image captioning without the need for expensive machine-translated data [130]. Instead of relying on large-scale multilingual datasets, LMCap introduces an image-blind few-shot multilingual captioning model. By prompting a language model with retrieved captions from a multilingual CLIP encoder, LMCap generates captions in the desired language using an XGLM decoder. This bypasses the need for multilingual caption data and achieves competitive performance with fully-supervised multilingual captioning models. Importantly, LMCap requires no supervised training on captioning data, making it efficient and effective, especially on geographically diverse datasets like XM3600.

Few-shot image captioning faces significant challenges, primarily due to models struggling to generalize from a small set of examples to diverse, unseen images, leading to overfitting. The relevance of the pre-trained models utilized for transfer learning and the caliber and diversity of the training instances are critical factors in determining the efficacy of the strategy. Complexity increases with the generation of nuanced language and accurate evaluation of it, and resources and experience are needed to fine-tune models on sparse data. Dependency on pre-trained models may also result in biases. To improve generalization and performance, overcoming these obstacles requires creative training approaches and model modifications.

Table 2 provides a comprehensive summary of the performance of notable methods for image captioning on the widely used COCO dataset. The evaluation is based on standard metrics, including BLEU (B-1 to B-4) for precision, METEOR (M) for harmonic mean, ROUGE-L (R) for recall-oriented understanding, and CIDEr (C) for consensus-based image description scoring. Fig. 9 illustrates the CIDEr scores for various image captioning methods on the COCO dataset. To complement this quantitative analysis, Fig. 10 showcases qualitative results, providing examples of captions generated by different models using the COCO dataset.

Table 2: An overview of notable recent studies on image captioning using COCO dataset

Method	B-1	B-2	B-3	B-4	M	R	C
RecallNet [54]	75.8	–	–	33.1	24.7	54.9	103.7
MRRC [131]	75.3	59.7	46.0	35.3	26.6	55.7	108.2
SCST [132]	–	–	–	34.2	26.7	55.7	114.0
HAN [133]	80.4	63.8	48.8	36.5	27.4	57.3	115.2
HAF [134]	80.5	62.9	47.7	35.5	27.3	–	116.4
TAAIC [79]	78.6	–	–	37.1	27.5	57.2	119.6
UpDown [52]	79.8	–	–	36.3	27.7	56.9	120.1
TDA+GLD [135]	78.8	62.6	48.0	36.1	27.8	57.1	121.1
RFNet [136]	79.1	63.1	48.4	36.5	27.7	57.3	121.9
Stack-VS [93]	79.4	63.6	49.0	37.2	27.9	57.7	122.6
VFDICM [137]	80.8	64.2	49.3	37.2	28.3	57.9	122.4
SDVFR [138]	81.1	64.3	49.5	37.4	28.4	58.1	123.2
ICEAP [139]	81.1	64.5	49.5	37.4	28.5	58.2	123.8
SGAE [140]	81.0	65.6	50.7	38.5	28.2	58.6	123.8
GCN-LSTM [141]	80.8	65.5	50.8	38.7	28.5	58.5	125.3
Dual Global [142]	80.8	65.6	51.1	39.1	28.9	58.9	126.3
AoA [143]	81.0	65.8	51.4	39.4	29.1	58.9	126.9
HIP [141]	81.6	66.2	51.5	39.3	28.8	59.0	127.9
M2Trans [144]	81.6	66.4	51.8	39.7	29.4	59.2	129.3
X-LAN [75]	81.9	66.9	52.4	40.3	29.6	59.5	131.1
DLCT [145]	82.4	67.4	52.8	40.6	29.8	59.8	133.3
GRIT [146]	84.1	69.4	54.9	42.5	30.9	61.2	141.3
CoCa [147]	–	–	–	40.9	33.9	–	143.6
SimVLM [148]	–	–	–	40.6	33.4	–	143.3
BLIP-2 ViT-G OPT 2.7B [149]	–	–	–	43.7	–	–	145.8
GIT [150]	–	–	–	44.1	32.2	–	151.1
OFA [151]	–	–	–	44.9	32.5	–	154.9
mPLUG [64]	–	–	–	46.5	32.0	–	155.1

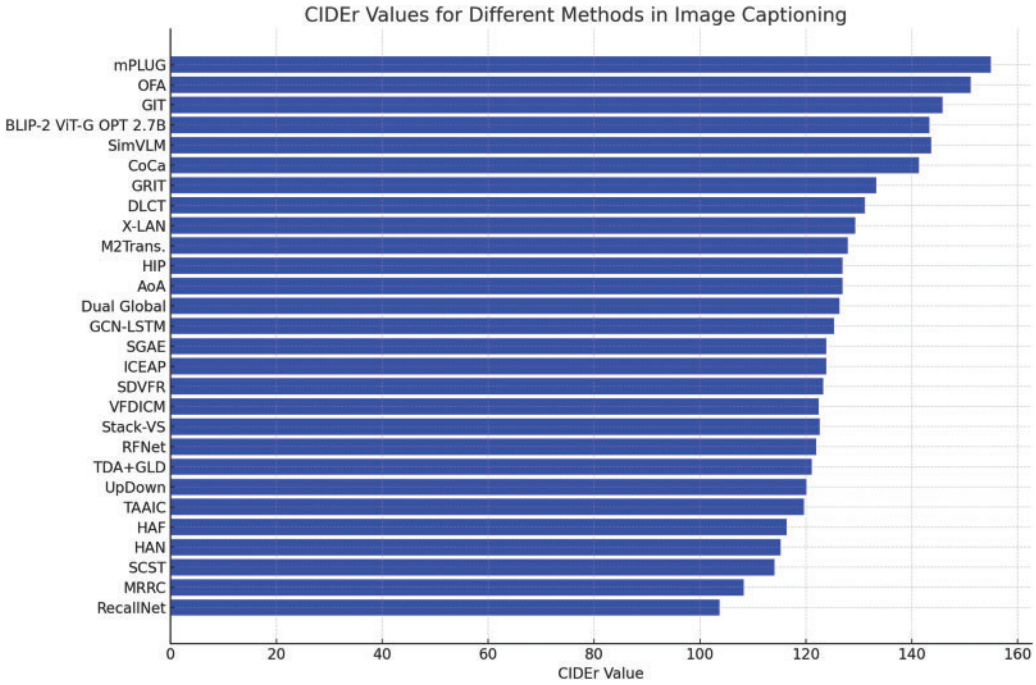

Figure 9: CIDEr scores for image captioning methods on the COCO dataset

B: a man wearing glasses and a tie in front of

N: a man in a suit and tie posing for a picture

S: a man wearing glasses and a black shirt and tie

GT1: A young man wearing black attire and a flowered tie is standing and smiling.

GT2: A man with glasses and his eyes closed dressed in a black shirt and a necktie.

GT3: A man in a green tie with his eyes closed.

GT4: Smiling man wearing black shirt and pale green tie.

GT5: A person that is dressed up very nicely.

B: a couple of people sitting at a table with food

N:a group of people sitting at a table with plates of food

S: a group of people sitting at a table eating pizza

GT1: A few men posing for a photo at a dinner table.

GT2: Group of men sitting at a dinner table with pizza on the table.

GT3: Three men eating a pizza meal at a table.

GT4: A group of people sitting around a wooden table.

GT5: Three men sitting at a table that has three pizzas on it.

Figure 10: (Continued)

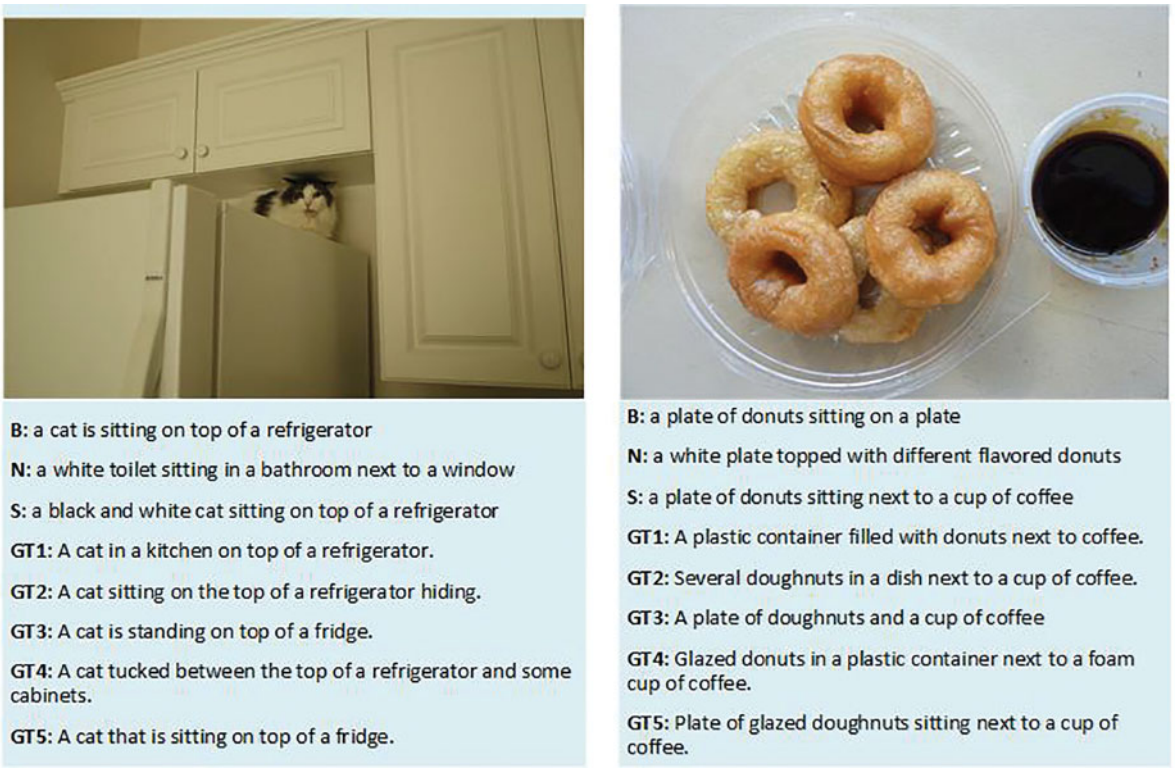

Figure 10: Examples of captions generated by different models using the COCO dataset are provided. Caption B is produced by the UpDown model, Caption N by the NumCap model, and captions S by the SDVFR model. GT refers to the ground truth captions [138]

4 Image Captioning Learning Approaches

Usually, captioning networks learn by predicting the subsequent word in reference captions, employing the traditional cross-entropy loss function. Evaluation of generated captions uses a suite of evaluation metrics. In recent times, there has been a surge in innovative approaches that directly fine-tune image captioning models with these evaluation metrics, even though they possess non-differentiable attributes. These optimization procedures leverage reinforcement learning methods [135,152] as a means to achieve this. One optimization strategy, self-critical sequence training (SCST), uses the CIDEr evaluation metric to enhance and train captioning models. SCST exhibits a substantial improvement in model performance, particularly in relation to the CIDEr metric [132].

Another innovative optimization method, the global-local discriminative objective, is based on reinforcement learning principles [135]. By introducing a global discriminative constraint, the model is encouraged to align the produced description with the corresponding image while avoiding similarity to other images. The local discriminative constraint, on the other hand, emphasizes less frequent but specific words. These constraints collectively contribute to generating more descriptive captions that capture visual details.

In a separate approach, an image annotating network built on the Transformer architecture introduced a modification to the maximum likelihood estimation (MLE) by incorporating a Kullback-Leibler (KL) divergence term. This term distinguishes between the model's prediction probability distribution and the ground truth distribution and aims to improve description generation [153]. Additionally, an approach that

leverages knowledge graphs enhances the Transformer network's description capabilities by considering word embeddings and their cosine similarity [153].

The hierarchical attention fusion (HAF) model serves as a reinforcement learning-based baseline for image captioning and incorporates multi-level feature maps from ResNet into hierarchical attention [134]. During the reinforcement learning phase, a revaluation network (REN) is used to reevaluate CIDEr scores and assign different weights to each word in the generated sentence based on word significance. This process is associated with word-level rewards. A scoring network (SN) also assesses the generated annotation against reference annotations, offering benefits from additional, unmatched references, which contributes to sentence-level rewards [134].

A novel reinforcement learning approach for image captioning, known as Vocabulary-Critical Sequence Training (VCST), introduces a vocabulary-critic module that assesses each word in the vocabulary at each generation step, providing different credits to each word. VCST incorporates efficient algorithms for BLEU and CIDEr-D metric calculations to minimize computation time costs, making it adaptable for integration into existing reinforcement learning approaches to enhance their performance [154].

In the context of language models, the initial masked language model, as introduced for BERT [155], aims to construct a comprehensive language representation by randomly concealing a portion of input tokens. It instructs the model to predict these concealed tokens while using adjacent tokens for context, allowing it to build a robust sentence representation. However, this training approach is slower compared to traditional left-to-right or right-to-left methods, as it focuses exclusively on predicting concealed tokens. Some research has adopted this method as a pre-training objective, sometimes without combining it with cross-entropy [126,156].

In the context of pre-training in vision and language, two primary methods are commonly used. The first involves masked contextual token loss, in which tokens from both textual and visual modalities are concealed, based on BERT [155]. This approach enhances the representation that combines both modalities. The second method is contrastive loss, in which inputs are organized into image regions, caption words, and object tag triples. The task is to differentiate between correct triples and those with randomly replaced tags [126,156]. There is also an objective focusing on aligning text and image within a word region, requiring the model to predict the original word sequence from corrupted word sequence [68].

In GPT-based image captioning, the learning approaches typically involve a combination of pre-training and fine-tuning techniques. Initially, masked language modeling and other unsupervised learning techniques are used to pre-train the GPT model on large text corpora [157]. The model gains a solid basis in natural language understanding during this pre-training phase. The model is fine-tuned on particular image-captioning datasets after pre-training. By combining text embeddings from the GPT model and picture attributes taken from pre-trained convolutional neural networks (CNNs) [158], the model learns to provide captions for images during fine-tuning. The alignment of textual and visual information is frequently facilitated by the use of attention processes [77].

5 Datasets and Evaluation Metrics

Image captioning models are developed using datasets and evaluation metrics. Datasets provide the foundation for training and testing these models. Evaluation metrics help us objectively assess caption quality and effectiveness. In this section, we delve into some of the prominent datasets and evaluation metrics, highlighting their significance and contributions to improving image captioning.

5.1 Standard Datasets

A variety of datasets have been introduced and extensively utilized in the field of image captioning. Prominent among these datasets are Flickr8K [159], Flickr30k [160], Microsoft COCO [161], and Visual Genome [162]. These datasets serve as crucial resources for researchers, providing rich and diverse sources of images and associated textual descriptions to facilitate advancements in image captioning research. Fig. 11 presents examples from the Flickr8K, Flickr30K, Microsoft COCO, and Visual Genome datasets.

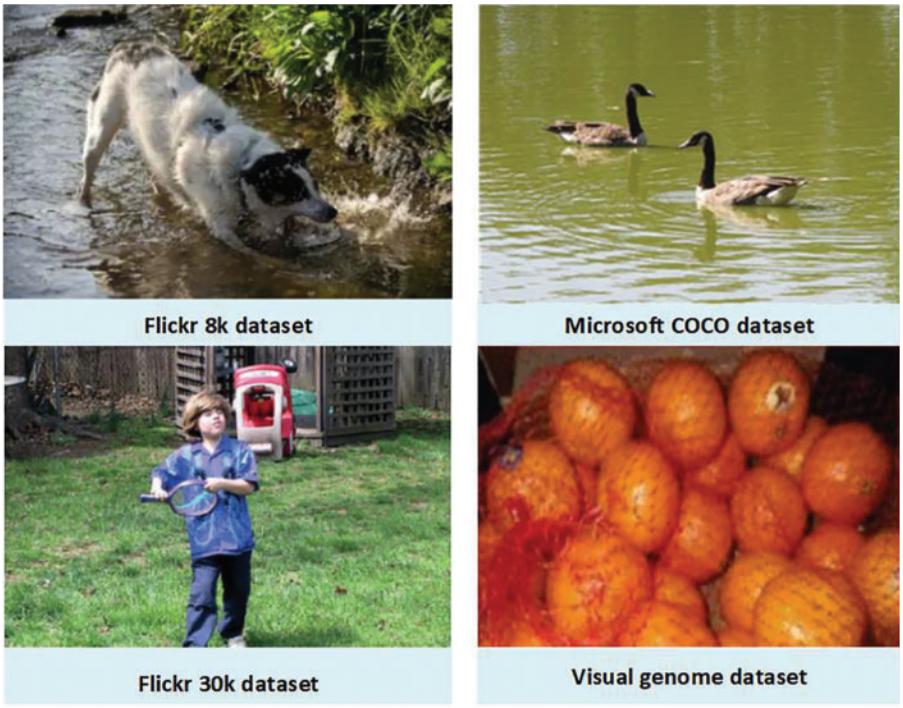

Figure 11: Examples from the Flickr8K, Flickr30K, Microsoft COCO, and visual genome datasets

- **Flickr8K** [159] dataset comprises around 8000 images sourced from diverse groups on Flickr.com, approximately 6000 images for training and 1000 images for each validation and testing [163]. It intentionally avoids focusing on famous locations or individuals, offering various everyday situations and locations. Notably, each image comes with five human-annotated captions, enriching the dataset's utility for image captioning models.
- **Flickr30K** [160], an extension of Flickr8K, offers a substantial collection of 31,783, typically around 29,000 images for training and 1000 images for each validation and testing set. Each image is associated with five descriptive sentences, yielding a wealth of textual data, and there are, on average, 16.6 expressions per image. This rich dataset proves invaluable for research in computer vision and natural language processing.
- **Microsoft COCO dataset** [161] is the largest and most widely recognized benchmark dataset in image captioning. It consists of 123,000 images, each accompanied by five ground-truth sentences. This dataset is commonly used to evaluate image captioning models, and its validation and testing sets contain 5000 images each, with 113,000 images in the training set. It serves as a standard reference for assessing the performance of image captioning techniques, according to Karpathy's data split [164], which is widely adopted in image captioning as in [99,165,166].

- **Visual genome dataset** [162] dataset stands out in image captioning due to its unique approach to considering the relationships between objects within images. Unlike some of its counterparts, the Visual Genome dataset does not provide captions for the entire scene; instead, it offers descriptions for different regions within an image, allowing for a more detailed and nuanced understanding. This rich dataset comprises over 100,000 images, each characterized by 18 attributes and featuring an average of 21 distinct objects. Furthermore, the dataset encompasses descriptions of 18 object relationships, making it a valuable resource for research in computer vision and language understanding. The Visual Genome dataset does not have predefined training and testing splits. Researchers create custom splits based on their specific research requirements, making the number of training and testing images variable.

5.2 Evaluation Metrics

Researchers have introduced a variety of evaluation metrics to provide quantitative estimates for the linguistic expressions generated from images. These metrics have become integral to virtually all image captioning research endeavors and are widely regarded as the standard benchmarks for assessment. Among these established evaluation metrics are BLEU [167], SPICE [168], CIDEr [169], METEOR [170], and ROUGE [171]. We offer a succinct overview of each of these widely recognized assessment criteria. BLEU, SPICE, CIDEr, METEOR, and ROUGE are quintessential evaluation metrics used to gauge the quality and accuracy of linguistic descriptions generated from images. These metrics play a pivotal role in image captioning research, offering a standardized framework for assessing and comparing the performance of various models. Below, we provide a concise introduction to each of these benchmark evaluation metrics.

- **BLEU** [167], short for Bilingual Evaluation Understudy, serves as an evaluation metric designed to assess the quality of machine-translated sentences. It involves comparing the generated translations to professional human translations, known as reference sentences, with the aim of gauging how closely the translated sentence matches the reference sentences. BLEU is recognized as one of the earliest evaluation metrics to quantify the proximity of a translation through a numerical value.
- **METEOR** [170], an evaluation metric employed for automated machine translation assessment, stands for Metric for Evaluation of Translation with Explicit Ordering. It relies on a broader concept of matching individual words (unigrams) between machine-generated translations and human-generated reference translations. It considers factors such as word meanings, stemmed forms, and surface forms for these unigram matches. Additionally, METEOR can be easily extended to incorporate more advanced matching techniques.
- **ROUGE** [171], short for Recall-Oriented Understudy for Gisting Evaluation, is a software tool employed to automatically assess summaries. It works by comparing machine-generated translations or summaries with reference translations or summaries. ROUGE calculates the number of overlapping units, encompassing word sequences, n-grams, and word pairs, between the human-crafted ideal summaries and the automatically generated summary.
- **CIDEr** [169], which stands for Consensus-based Image Description Evaluation, is an evaluation metric tailored for the assessment of image annotation tasks. Essentially, CIDEr gauges the agreement between the set of human-written reference annotations for an image and the description generated for that image. To accomplish this, stemming is employed to reduce reference and generate caption words in their root forms. Additionally, CIDEr utilizes Term Frequency Inverse Document Frequency (TF-IDF) to assign weights to individual n-grams.
- **SPICE** [168], short for Semantic Propositional Image Captioning Evaluation, is a metric employed for the evaluation of image and video annotations. Its primary function is to determine the similarity between the scene graph tuple extracted from the human-authored reference annotations and the annotations

generated by a computer. The semantic scene graph, created through a dependency parse tree, encodes information about objects, their relationships, and their attributes.

6 Emerging Challenges and Future Directions

6.1 Current Challenges

- Dataset Constraints: The limitations of existing datasets are a significant hurdle in image captioning. Many datasets lack diversity, leading to biased models that produce repetitive and less natural captions. For example, in medical imaging datasets, the imbalance between diseased and normal samples often results in captions that default to normalcy. Enhancements in image and text alignment algorithms, advanced unsupervised learning methods, and more comprehensive paired datasets are necessary to address these issues [34,172].
- Ethical and Bias Considerations: Several real-world deployments and user studies highlight both the promise and challenges of image captioning in practical applications. For instance, Microsoft's Seeing AI app assists visually impaired users by narrating their surroundings, with iterative improvements based on user feedback to enhance contextual accuracy [173]. Similarly, in autonomous driving, systems like Tesla's Autopilot and Waymo's self-driving vehicles leverage image-based scene interpretation, showing success in controlled settings but facing limitations in complex or low-visibility scenarios [174]. Social media platforms, including Facebook and Instagram, also utilize captioning to improve accessibility, automatically generating descriptions that help visually impaired users interact with image-based content [175]. Meanwhile, medical imaging applications use captioning models to support radiologists by providing preliminary descriptions, though human oversight remains essential for accuracy and safety.
- Model Design: Advanced deep learning frameworks and reinforcement learning techniques hold potential to enhance captioning performance. However, designing effective reward structures and ensuring continuous learning without instability remains challenging. For instance, Generative Adversarial Networks (GANs) frequently struggle with convergence, highlighting the need for more robust training strategies and simpler model architectures [176,177].
- Interpretability Issues: Explaining and understanding the process behind image captioning is crucial. Current models are often evaluated using metrics from machine translation or natural language processing, which may not align with human cognitive processes. Developing evaluation methods that truly reflect human cognitive processes is essential for improving the interpretability of image captioning models [35,62].
- Real-time and Interactive Captioning: Developing models capable of generating real-time and interactive captions remains a challenge. These models must handle dynamic and complex scenes, requiring advanced algorithms and optimized processing capabilities to provide timely and relevant captions [35,178].
- Multimodal Integration: Effective integration of visual and textual data is critical for generating meaningful captions. Current models often struggle with aligning these modalities, particularly in unstructured and noisy environments. Enhancing multimodal understanding and improving the robustness of models in diverse settings are ongoing challenges [179].

6.2 Future Directions

- Unsupervised Learning Approaches: Due to the limitations of supervised methods, there is growing interest in unsupervised techniques. These methods aim to mitigate dataset bias and the object-hallucination phenomenon seen in supervised learning. However, the challenge of aligning information from different modalities in an unpaired setting remains significant [34,35,62].

- Vision-Language Pre-training (VLP): VLP methods have shown potential in addressing some limitations of supervised methods and object detector-based designs. Detector-free designs, which utilize general visual encoders for cross-modal fusion, require further investigation to enhance their capabilities for generative tasks like image captioning [180].
- Adaptive Learning and Personalization: Future models may incorporate adaptive learning techniques to personalize caption generation based on user preferences and contextual requirements. This involves developing algorithms that can learn and adapt over time, providing more tailored and relevant captions for different users and scenarios [35,179].
- Applications for the Visually Impaired: There is increasing focus on using image captioning technology to assist the visually impaired. Developing vision assistants that can accurately describe images and inform users about their surroundings can significantly improve the quality of life for visually impaired individuals [62,178].
- Ethical Considerations and Evaluation Metrics: Addressing ethical biases in caption generation, improving real-time and interactive captioning capabilities, and refining evaluation metrics are crucial areas of ongoing research. Ensuring that models do not reinforce harmful stereotypes is essential for the responsible advancement of image captioning technology [178,179].
- Transformer Models and Attention Mechanisms: Recent advancements in transformer models and attention mechanisms have shown significant promise in improving image captioning. Designing novel attention technologies that mimic human cognitive processes can further enhance the extraction of representative and targeted features, leading to more accurate and contextually relevant captions [35].
- Utilization of Scene Graphs: Scene graphs offer a promising approach for image captioning by representing complex relationships and interactions between objects. Improving scene graph parsers and pre-training with visually relevant relational data are critical areas for future research [62].

By addressing these challenges and embracing emerging trends, researchers can unlock the full potential of image annotation systems, making them more accurate, reliable, and applicable to a broader range of real-world scenarios.

7 Industry Impact of Advanced Image Captioning

The advancements in image captioning discussed in this paper hold significant implications for practitioners, particularly in industry settings. Enhanced captioning accuracy and contextual relevance can directly benefit sectors like e-commerce, media, and accessibility services, providing better support for visually impaired individuals and improving content discoverability. Workflow automation in fields such as customer support and medical diagnostics can also benefit from these refined models, which reduce manual intervention and errors by automating image tagging, content categorization, and medical image reporting. In marketing and content generation, dynamic captioning models enable personalization based on user interaction data, making images more engaging, especially in news media contexts. Furthermore, improved captioning accuracy aids autonomous systems in interpreting complex environments, crucial for the safety and compliance of applications like autonomous vehicles and robotics. Finally, cutting-edge captioning technology enhances scalability and cost-efficiency for large-scale applications, such as social media platforms, by reducing reliance on extensive human resources and efficiently managing vast amounts of data. Overall, these advancements underscore the transformative potential of image captioning in improving operational efficiency, safety, and user engagement across various industries.

8 Conclusions

Image captioning, the process of generating textual descriptions for visual content, has significant implications across various disciplines. It facilitates accessibility for individuals with visual impairments, supports autonomous navigation in self-driving vehicles, streamlines medical image analysis, and enhances news articles' informativeness. The field of image captioning has witnessed substantial evolution, with diverse methodologies and models, from traditional encoder-decoder frameworks to advanced transformer-based architectures, consistently improving its capabilities. This survey endeavors to provide a comprehensive exploration of image captioning strategies and techniques aimed at enhancing the generation of image captions, encompassing image captioning with visual attention, image captioning with semantic information, multi-caption image captioning, neural architecture search, few-shot image, multi-phase learning, and cross-modal embedding, as well as image captioning learning approaches, datasets, and evaluation metrics. Future research in image captioning has promising potential in several areas. Firstly, exploring methods to integrate contextual information into models could enhance the relevance of generated captions. Improving dataset handling techniques like data augmentation and cleaning can lead to more robust model training. Leveraging multimodal data sources such as text and audio can enrich the understanding of visual content. Novel approaches like reinforcement learning and self-supervised learning may enhance caption quality and creativity. Standardizing evaluation metrics will enable fair comparisons and facilitate advancements in the field.

Acknowledgement: The authors would like to thanks Prince Sultan University for their support.

Funding Statement: This work is supported by the National Natural Science Foundation of China (Nos. U22A2034, 62177047), High Caliber Foreign Experts Introduction Plan funded by MOST, and Central South University Research Programme of Advanced Interdisciplinary Studies (No. 2023QYJC020). Also, the authors would like to thank Prince Sultan University for paying the APC of this article.

Author Contributions: The authors' contributions to this work are as follows: Alaa Thobhani led the investigation, conceptualization, visualization, and software development. Xiaoyan Kui assisted with the manuscript review and editing. Beiji Zou supervised the entire project. Amr Abdussalam participated in the investigation, while Muhammad Asim provided resources. Sajid Shah conducted the formal analysis, and Mohammed ELAffendi was responsible for validation. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Not applicable.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflict of interests regarding the publication of this article.

References

1. Alkhonaini MA, Mengash HA, Nemri N, Ebad SA, Alotaibi FA, Aljabri J, et al. From pixels to predictions: role of boosted deep learning enabled object detection for autonomous vehicles on large scale consumer electronics environment. Fractals. 2024;32:1–17. doi:10.1142/S0218348X2540047X.
2. Ayesha J, Shahid N, Li J, Tariq M, Kashif J, Amjad R, et al. Deep transfer learning-based automated diabetic retinopathy detection using retinal fundus images in remote areas. Int J Comput Intell Syst. 2024;17(1):135. doi:10.1007/s44196-024-00520-w.
3. Safa RW, Norhaida MS, Mohd SMR, Amjad RK, Saeed AB, Tanzila S. Synergistic integration of transfer learning and deep learning for enhanced object detection in digital images. IEEE Access. 2024;12(4):13525–36. doi:10.1109/ACCESS.2024.3354706.

4. Honda U, Watanabe T, Matsumoto Y. Switching to discriminative image captioning by relieving a bottleneck of reinforcement learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision; 2023. p. 1124–34.
5. Cho S, Oh H. Generalized image captioning for multilingual support. Appl Sci. 2023;13(4):2446. doi:10.3390/app13042446.
6. Gui J, Sun Z, Wen Y, Tao D, Ye J. A review on generative adversarial networks: algorithms, theory, and applications. IEEE Transact Knowl Data Eng. 2021;35(4):3313–32. doi:10.1109/TKDE.2021.3130191.
7. Yang Y, Yu J, Zhang J, Han W, Jiang H, Huang Q. Joint embedding of deep visual and semantic features for medical image report generation. IEEE Transact Multimed. 2021;25(107):167–78. doi:10.1109/TMM.2021.3122542.
8. Ayesha H, Iqbal S, Tariq M, Abrar M, Sanaullah M, Abbas I, et al. Automatic medical image interpretation: state of the art and future directions. Pattern Recognit. 2021;114:107856. doi:10.1016/j.patcog.2021.107856.
9. Liu F, Wang Y, Wang T, Ordonez V. Visual news: benchmark and challenges in news image captioning. arXiv:201003743. 2020.
10. Tran A, Mathews A, Xie L. Transform and tell: entity-aware news image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020. p. 13035–45.
11. Yang X, Karaman S, Tetreault J, Jaimes A. Journalistic guidelines aware news image captioning. arXiv:210902865. 2021.
12. Szafir D, Szafir DA. Connecting human-robot interaction and data visualization. In: Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction; 2021. p. 281–92.
13. Sudhakar J, Iyer VV, Sharmila ST. Image caption generation using deep neural networks. In: 2022 International Conference for Advancement in Technology (ICONAT); 2022; IEEE. p. 1–3.
14. Li Z, Mu Y, Sun Z, Song S, Su J, Zhang J. Intention understanding in human–robot interaction based on visual-NLP semantics. Front Neurorobot. 2021;14:610139. doi:10.3389/fnbot.2020.610139.
15. Das S, Jain L, Das A. Deep learning for military image captioning. In: 2018 21st International Conference on Information Fusion (FUSION); 2018; IEEE. p. 2165–71.
16. Ghataoura D, Ogbonnaya S. Application of image captioning and retrieval to support military decision making. In: 2021 International Conference on Military Communication and Information Systems (ICMCIS); 2021; IEEE. p. 1–8.
17. Ma X, Zhao R, Shi Z. Multiscale methods for optical remote-sensing image captioning. IEEE Geosci Remote Sens Lett. 2020;18(11):2001–5. doi:10.1109/LGRS.2020.3009243.
18. Sharma H, Agrahari M, Singh SK, Firoj M, Mishra RK. Image captioning: a comprehensive survey. In: 2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC); 2020; IEEE. p. 325–8.
19. Nursikuwagus A, Munir R, Khodra ML. Hybrid of deep learning and word embedding in generating captions: image-captioning solution for geological rock images. J Imaging. 2022;8(11):294. doi:10.3390/jimaging8110294.
20. Song K, Chen L, Wang H. Style-enhanced transformer for image captioning in construction scenes. Entropy. 2024;26(3):224. doi:10.3390/e26030224.
21. Amirian S, Rasheed K, Taha TR, Arabnia HR. A short review on image caption generation with deep learning. In: Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV); 2019. p. 10–8.
22. Amirian S, Rasheed K, Taha TR, Arabnia HR. Image captioning with generative adversarial network. In: 2019 International Conference on Computational Science and Computational Intelligence (CSCI); 2019; IEEE. p. 272–5.
23. Aneja J, Deshpande A, Schwing AG. Convolutional image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018. p. 5561–70.
24. Staniūtė R, Šešok D. A systematic literature review on image captioning. Appl Sci. 2019;9(10):2024. doi:10.3390/app9102024.
25. Bernardi R, Cakici R, Elliott D, Erdem A, Erdem E, Ikizler-Cinbis N, et al. Automatic description generation from images: a survey of models, datasets, and evaluation measures. J Artifl Intell Res. 2016;55:409–42. doi:10.1613/jair.4900.

26. Kumar A, Goel S. A survey of evolution of image captioning techniques. Int J Hybrid Intell Syst. 2017;14(3):123–39. doi:10.3233/HIS-170246.
27. Bai S, An S. A survey on automatic image caption generation. Neurocomputing. 2018;311(1):291–304. doi:10.1016/j.neucom.2018.05.080.
28. Li S, Tao Z, Li K, Fu Y. Visual to text: survey of image and video captioning. IEEE Transact Emerg Top Computat Intell. 2019;3(4):297–312. doi:10.1109/TETCI.2019.2892755.
29. Amirian S, Rasheed K, Taha TR, Arabnia HR. Automatic image and video caption generation with deep learning: a concise review and algorithmic overlap. IEEE Access. 2020;8:218386–400. doi:10.1109/ACCESS.2020.3042484.
30. Alam M, Samad MD, Vidyaratne L, Glandon A, Iftekharuddin KM. Survey on deep neural networks in speech and vision systems. Neurocomputing. 2020;417(2):302–21. doi:10.1016/j.neucom.2020.07.053.
31. Luo G, Cheng L, Jing C, Zhao C, Song G. A thorough review of models, evaluation metrics, and datasets on image captioning. IET Image Process. 2022;16(2):311–32. doi:10.1049/ipr2.12367.
32. Zohourianshahzadi Z, Kalita JK. Neural attention for image captioning: review of outstanding methods. Artif Intell Rev. 2022;55(5):3833–62. doi:10.1007/s10462-021-10092-2.
33. Elhagry A, Kadaoui K. A thorough review on recent deep learning methodologies for image captioning. arXiv:210713114. 2021.
34. Sharma H, Padha D. A comprehensive survey on image captioning: from handcrafted to deep learning-based techniques, a taxonomy and open research issues. Artif Intell Rev. 2023;56(11):1–43. doi:10.1007/s10462-023-10488-2.
35. Xu L, Tang Q, Lv J, Zheng B, Zeng X, Li W. Deep image captioning: a review of methods, trends and future challenges. Neurocomputing. 2023;546(10):126287. doi:10.1016/j.neucom.2023.126287.
36. Yang L, Zhang Z, Song Y, Hong S, Xu R, Zhao Y, et al. Diffusion models: a comprehensive survey of methods and applications. ACM Comput Surv. 2022;56(4):105. doi:10.1145/362623.
37. Yumeng Z, Jing Y, Shuo G, Limin L. News image-text matching with news knowledge graph. IEEE Access. 2021;9:108017–27. doi:10.1109/ACCESS.2021.3093650.
38. Daneshfar F, Bartani A, Lotfi P. Image captioning by diffusion models: a survey. Eng Appl Artif Intell. 2024;138(1):109288. doi:10.1016/j.engappai.2024.109288.
39. Daneshfar F, Saifee BS, Soleymanbaigi S, Aeini M. Elastic deep multi-view autoencoder with diversity embedding. Inf Sci. 2025;689:121482. doi:10.1016/j.ins.2024.121482.
40. Jian M, Wang J, Yu H, Wang GG. Integrating object proposal with attention networks for video saliency detection. Informat Sci. 2021;576(12):819–30. doi:10.1016/j.ins.2021.08.069.
41. Jian M, Wang J, Yu H, Wang G, Meng X, Yang L, et al. Visual saliency detection by integrating spatial position prior of object with background cues. Expert Syst Appl. 2021;168(11):114219. doi:10.1016/j.eswa.2020.114219.
42. Fan DP, Ji GP, Cheng MM, Shao L. Concealed object detection. IEEE Transact Pattern Anal Mach Intell. 2021;44(10):6024–42. doi:10.1109/TPAMI.2021.3085766.
43. Feng D, Wang Z, Zhou Y, Rosenbaum L, Timm F, Dietmayer K, et al. Labels are not perfect: inferring spatial uncertainty in object detection. IEEE Transact Intell Transport Syst. 2021;23(8):9981–94. doi:10.1109/TITS.2021.3096943.
44. Huang J, Yan W, Li T, Liu S, Li G. Learning the global descriptor for 3-D object recognition based on multiple views decomposition. IEEE Transact Multim. 2020;24:188–201. doi:10.1109/TMM.2020.3047762.
45. Tian Y, Song W, Chen L, Fong S, Sung Y, Kwak J. A 3D object recognition method from LiDAR point cloud based on USAE-BLS. IEEE Transact Intell Transport Syst. 2022;23(9):15267–77. doi:10.1109/TITS.2021.3140112.
46. Fu J, Rui Y. Advances in deep learning approaches for image tagging. APSIPA Trans Signal Inf Process. 2017;6(1):e11. doi:10.1017/ATSIP.2017.12.
47. Cui C, Shen J, Ma J, Lian T. Social tag relevance learning via ranking-oriented neighbor voting. Multimed Tools Appl. 2017;76(6):8831–57. doi:10.1007/s11042-016-3512-1.
48. Tang J, Chen Q, Wang M, Yan S, Chua TS, Jain R. Towards optimizing human labeling for interactive image tagging. ACM Transact Multime Comput Commun Appl. 2013;9(4):1–18. doi:10.1145/2501643.2501651.

49. Shen J, Wang M, Yan S, Hua XS. Multimedia tagging: past, present and future. In: Proceedings of the 19th ACM International Conference on Multimedia; 2011. p. 639–40.
50. Wang XJ, Zhang L, Liu M, Li Y, Ma WY. Arista-image search to annotation on billions of web photos. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition; 2010; IEEE. p. 2987–94.
51. Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: a neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2015. p. 3156–64.
52. Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, et al. Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018. p. 6077–86.
53. Xiao X, Wang L, Ding K, Xiang S, Pan C. Deep hierarchical encoder–decoder network for image captioning. IEEE Transact Multim. 2019;21(11):2942–56. doi:10.1109/TMM.2019.2915033.
54. Wu L, Xu M, Wang J, Perry S. Recall what you see continually using GridLSTM in image captioning. IEEE Transact Multim. 2019;22(3):808–18. doi:10.1109/TMM.2019.2931815.
55. Jia X, Gavves E, Fernando B, Tuytelaars T. Guiding the long-short term memory model for image caption generation. In: Proceedings of the IEEE International Conference on Computer Vision; 2015. p. 2407–15.
56. Chen X, Lawrence Zitnick C. Mind's eye: a recurrent visual representation for image caption generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2015. p. 2422–31.
57. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition, CoRR abs/1409.1556. arXiv:14091556. 2015.
58. You Q, Jin H, Wang Z, Fang C, Luo J. Image captioning with semantic attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016. p. 4651–9.
59. Gu J, Wang G, Cai J, Chen T. An empirical study of language CNN for image captioning. In: Proceedings of the IEEE International Conference on Computer Vision; 2017. p. 1222–31.
60. Chen F, Ji R, Su J, Wu Y, Wu Y. Structcap: structured semantic embedding for image captioning. In: Proceedings of the 25th ACM International Conference on Multimedia; 2017. p. 46–54.
61. Chen F, Ji R, Sun X, Wu Y, Su J. Groupcap: group-based image captioning with structured relevance and diversity constraints. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018. p. 1345–53.
62. Stefanini M, Cornia M, Baraldi L, Cascianelli S, Fiameni G, Cucchiara R. From show to tell: a survey on deep learning-based image captioning. IEEE Transact Pattern Anal Mach Intell. 2022;45(1):539–59. doi:10.1109/TPAMI.2022.3148210.
63. Zhou L, Palangi H, Zhang L, Hu H, Corso J, Gao J. Unified vision-language pre-training for image captioning and VQA. Proc AAAI Conf Artif Intell. 2020;34(07):13041–9. doi:10.1609/aaai.v34i07.7005.
64. Li C, Xu H, Tian J, Wang W, Yan M, Bi B, et al. mPLUG: effective and efficient vision-language learning by cross-modal skip-connections. arXiv:220512005. 2022.
65. Liu S, Fan L, Johns E, Yu Z, Xiao C, Anandkumar A. Prismer: a vision-language model with an ensemble of experts. arXiv:230302506. 2023.
66. Bai J, Bai S, Yang S, Wang S, Tan S, Wang P, et al. Qwen-VL: a versatile vision-language model for understanding. arXiv:2308.12966. 2023.
67. Chen J, Guo H, Yi K, Li B, Elhoseiny M. Visualgpt: data-efficient image captioning by balancing visual input and linguistic knowledge from pretraining. arXiv:210210407. 2021.
68. Xia Q, Huang H, Duan N, Zhang D, Ji L, Sui Z, et al. XGPT: cross-modal generative pre-training for image captioning. In: Natural Language Processing and Chinese Computing: 10th CCF International Conference, NLPCC; 2021 Oct 13–17; Qingdao, China: Springer; 2021. p. 786–97.
69. Mokady R, Hertz A, Bermano AH. Clipcap: clip prefix for image captioning. arXiv:211109734. 2021.
70. Luo Z, Xi Y, Zhang R, Ma J. A frustratingly simple approach for end-to-end image captioning. arXiv:220112723. 2022.
71. Hu Y, Hua H, Yang Z, Shi W, Smith NA, Luo J. PromptCap: prompt-guided image captioning for VQA with GPT-3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2023. p. 2963–75.

72. Ramos R, Martins B, Elliott D, Kementchedjhieva Y. Smallcap: lightweight image captioning prompted with retrieval augmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023. p. 2840–9.
73. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, et al. Show, attend and tell: neural image caption generation with visual attention. In: International Conference on Machine Learning; 2015; PMLR. p. 2048–57.
74. Jiang W, Wang W, Hu H. Bi-directional co-attention network for image captioning. ACM Transact Multimedia Comput Commun Appl. 2021;17(4):1–20. doi:10.1145/3460474.
75. Pan Y, Yao T, Li Y, Mei T. X-linear attention networks for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020. p. 10971–80.
76. Gao L, Li X, Song J, Shen HT. Hierarchical LSTMs with adaptive attention for visual captioning. IEEE Transact Pattern Anal Mach Intell. 2019;42(5):1112–31. doi:10.1109/TPAMI.2019.2894139.
77. Lu J, Xiong C, Parikh D, Socher R. 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; 2017. p. 375–83.
78. Ji J, Xu C, Zhang X, Wang B, Song X. Spatio-temporal memory attention for image captioning. IEEE Transact Image Process. 2020;29:7615–28. doi:10.1109/TIP.2020.3004729.
79. Yan C, Hao Y, Li L, Yin J, Liu A, Mao Z, et al. Task-adaptive attention for image captioning. IEEE Transact Circ Syst Video Technol. 2021;32(1):43–51. doi:10.1109/TCSVT.2021.3067449.
80. Yu L, Zhang J, Wu Q. Dual attention on pyramid feature maps for image captioning. IEEE Transact Multim. 2021;24:1775–86. doi:10.1109/TMM.2021.3072479.
81. Jiang W, Zhu M, Fang Y, Shi G, Zhao X, Liu Y. Visual cluster grounding for image captioning. IEEE Transact Image Process. 2022;31:3920–34. doi:10.1109/TIP.2022.3177318.
82. Yu J, Li J, Yu Z, Huang Q. Multimodal transformer with multi-view visual representation for image captioning. IEEE Transact Circ Syst Video Technol. 2019;30(12):4467–80. doi:10.1109/TCSVT.2019.2947482.
83. Liu AA, Zhai Y, Xu N, Nie W, Li W, Zhang Y. Region-aware image captioning via interaction learning. IEEE Transact Circ Syst Video Technol. 2021;32(6):3685–96. doi:10.1109/TCSVT.2021.3107035.
84. Guo MH, Lu CZ, Liu ZN, Cheng MM, Hu SM. Visual attention network. Computat Visual Media. 2023;9(4):733–52. doi:10.1007/s41095-023-0364-2.
85. Mehrani P, Tsotsos JK. Self-attention in vision transformers performs perceptual grouping, not attention. arXiv:230301542. 2023.
86. Shi B, Darrell T, Wang X. Top-down visual attention from analysis by synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023. p. 2102–12.
87. Al-Qatf M, Hawbani A, Wang X, Abdusallam A, Alsamhi S, Alhabib M, et al. RVAIC: refined visual attention for improved image captioning. J Intell Fuzzy Syst. 2024;46(2):3447–59. doi:10.3233/JIFS-233004.
88. Hossen MB, Ye Z, Abdussalam A, Hossain MI. GVA: guided visual attention approach for automatic image caption generation. Multimed Syst. 2024;30(1):50. doi:10.1007/s00530-023-01249-w.
89. Zhang M, Yang Y, Zhang H, Ji Y, Shen HT, Chua TS. More is better: precise and detailed image captioning using online positive recall and missing concepts mining. IEEE Transact Image Process. 2018;28(1):32–44. doi:10.1109/TIP.2018.2855415.
90. Huang Y, Chen J, Ouyang W, Wan W, Xue Y. Image captioning with end-to-end attribute detection and subsequent attributes prediction. IEEE Transact Image Process. 2020;29:4013–26. doi:10.1109/TIP.2020.2969330.
91. El-Gayar M. Automatic generation of image caption based on semantic relation using deep visual attention prediction. Int J Adv Comput Sci Appl. 2023;14(9). doi:10.14569/issn.2156-5570.
92. Wu Q, Shen C, Wang P, Dick A, Van Den Hengel A. Image captioning and visual question answering based on attributes and external knowledge. IEEE Transact Pattern Anal Mach Intell. 2017;40(6):1367–81. doi:10.1109/TPAMI.2017.2708709.
93. Cheng L, Wei W, Mao X, Liu Y, Miao C. Stack-VS: stacked visual-semantic attention for image caption generation. IEEE Access. 2020;8:154953–65. doi:10.1109/ACCESS.2020.3018752.
94. Rotstein N, Bensaid D, Brody S, Ganz R, Kimmel R. FuseCap: leveraging large language models to fuse visual data into enriched image captions. arXiv:230517718. 2023.

95. Haque N, Labiba I, Akter S. FaceAtt: enhancing image captioning with facial attributes for portrait images. arXiv:230913601. 2023.
96. Wang C, Gu X. Learning joint relationship attention network for image captioning. Expert Syst Appl. 2023;211(20):118474. doi:10.1016/j.eswa.2022.118474.
97. Hossen MB, Ye Z, Abdussalam A, Hassan SU. Attribute-driven filtering: a new attributes predicting approach for fine-grained image captioning. Eng Appl Artif Intell. 2024;137(4):109134. doi:10.1016/j.engappai.2024.109134.
98. Bae JW, Lee SH, Kim WY, Seong JH, Seo DH. Image captioning model using part-of-speech guidance module for description with diverse vocabulary. IEEE Access. 2022;10(11):45219–29. doi:10.1109/ACCESS.2022.3169781.
99. Zhang J, Mei K, Zheng Y, Fan J. Integrating part of speech guidance for image captioning. IEEE Transact Multim. 2020;23:92–104. doi:10.1109/TMM.2020.2976552.
100. Al-Qatf M, Hawbani A, Wang X, Abdusallam A, Zhao L, Alsamhi SH, et al. NPoSC-A3: a novel part of speech clues-aware adaptive attention mechanism for image captioning. Eng Appl Artif Intell. 2024;131(4):107732. doi:10.1016/j.engappai.2023.107732.
101. Yu N, Hu X, Song B, Yang J, Zhang J. Topic-oriented image captioning based on order-embedding. IEEE Transact Image Process. 2018;28(6):2743–54. doi:10.1109/TIP.2018.2889922.
102. Zhu Z, Xue Z, Yuan Z. Topic-guided attention for image captioning. In: 2018 25th IEEE International Conference on Image Processing (ICIP); 2018; IEEE. p. 2615–9.
103. Chen F, Xie S, Li X, Li S, Tang J, Wang T. What topics do images say: a neural image captioning model with topic representation. In: 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW); 2019; IEEE. p. 447–52.
104. Zia U, Riaz MM, Ghafoor A, Ali SS. Topic sensitive image descriptions. Neural Comput Appl. 2020;32(14):10471–9. doi:10.1007/s00521-019-04587-x.
105. Dash SK, Acharya S, Pakray P, Das R, Gelbukh A. Topic-based image caption generation. Arabian J Sci Eng. 2020;45(4):3025–34. doi:10.1007/s13369-019-04262-2.
106. Al-Qatf M, Wang X, Hawbani A, Abdussalam A, Alsamhi SH. Image captioning with novel topics guidance and retrieval-based topics re-weighting. IEEE Transact Multim. 2023;25:5984–99. doi:10.1109/TMM.2022.3202690.
107. Zhou L, Zhang Y, Jiang YG, Zhang T, Fan W. Re-caption: saliency-enhanced image captioning through two-phase learning. IEEE Transact Image Process. 2019;29:694–709. doi:10.1109/TIP.2019.2928144.
108. Ahmad S, Haq MU, Sethi MAJ, El Affendi MA, Farid Z, Al Luhaidan AS. Mapping faces from above: exploring face recognition algorithms and datasets for aerial drone images. In: Deep cognitive modelling in remote sensing image processing. IGI Global; 2024. p. 55–69.
109. Dai B, Fidler S, Urtasun R, Lin D. Towards diverse and natural image descriptions via a conditional GAN. In: Proceedings of the IEEE International Conference on Computer Vision; 2017. p. 2970–9.
110. Mao Y, Zhou C, Wang X, Li R. Show and tell more: topic-oriented multi-sentence image captioning. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18); 2018. p. 4258–64.
111. Xu G, Niu S, Tan M, Luo Y, Du Q, Wu Q. Towards accurate text-based image captioning with content diversity exploration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021. p. 12637–46.
112. Abdussalam A, Ye Z, Hawbani A, Al-Qatf M, Khan R. NumCap: a number-controlled multi-caption image captioning network. ACM Transact Multim Comput, Commun Appl. 2023;19(4):1–24. doi:10.1145/3576927.
113. Zhu X, Wang W, Guo L, Liu J. AutoCaption: image captioning with neural architecture search. arXiv:201209742. 2020.
114. Yu Z, Cui Y, Yu J, Wang M, Tao D, Tian Q. Deep multimodal neural architecture search. In: Proceedings of the 28th ACM International Conference on Multimedia; 2020. p. 3743–52.
115. Hosseini R, Xie P. Image understanding by captioning with differentiable architecture search. In: Proceedings of the 30th ACM International Conference on Multimedia; 2022. p. 4665–73.
116. Ling H, Fidler S. Teaching machines to describe images with natural language feedback. Adv Neural Inf Process Syst. 2017;30.

117. Gu J, Cai J, Wang G, Chen T. Stack-captioning: coarse-to-fine learning for image captioning. In: Proceedings of the AAAI Conference on Artificial Intelligence; 2018. Vol. 32.
118. Dong X, Zhu L, Zhang D, Yang Y, Wu F. Fast parameter adaptation for few-shot image captioning and visual question answering. In: Proceedings of the 26th ACM International Conference on Multimedia; 2018. p. 54–62.
119. Tsimpoukelli M, Menick JL, Cabi S, Eslami S, Vinyals O, Hill F. Multimodal few-shot learning with frozen language models. Adv Neural Inform Process Syst. 2021;34:200–12.
120. Yang Z, Ping W, Liu Z, Korthikanti V, Nie W, Huang DA, et al. Re-ViLM: retrieval-augmented visual language model for zero and few-shot image captioning. arXiv:230204858. 2023.
121. Faghri F, Fleet DJ, Kiros JR, Fidler S. VSE++: improving visual-semantic embeddings with hard negatives. arXiv:170705612. 2017.
122. Kiros R, Salakhutdinov R, Zemel R. Multimodal neural language models. In: International Conference on Machine Learning; 2014; PMLR. p. 595–603.
123. Vendrov I, Kiros R, Fidler S, Urtasun R. Order-embeddings of images and language. arXiv:151106361. 2015.
124. Dutta T, Biswas S. Generalized zero-shot cross-modal retrieval. IEEE Transact Image Process. 2019;28(12):5953–62. doi:10.1109/TIP.2019.2923287.
125. Wang S, Guo D, Xu X, Zhuo L, Wang M. Cross-modality retrieval by joint correlation learning. ACM Transact Multim Comput Commun Appl. 2019;15(2s):1–16. doi:10.1145/3314577.
126. Li X, Yin X, Li C, Zhang P, Hu X, Zhang L, et al. Oscar: object-semantics aligned pre-training for vision-language tasks. In: Computer Vision–ECCV 2020: 16th European Conference; 2020 Aug 23–28; Glasgow, UK: Springer. p. 121–37.
127. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning; 2021; PMLR. p. 8748–63.
128. Chen X, Jiang M, Zhao Q. Self-distillation for few-shot image captioning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision; 2021. p. 545–55.
129. Yang X, Wu Y, Yang M, Chen H, Geng X. Exploring diverse in-context configurations for image captioning. Adv Neural Inf Process Syst. 2024;36.
130. Ramos R, Martins B, Elliott D. LMCap: few-shot multilingual image captioning by retrieval augmented language model prompting. arXiv:230519821. 2023.
131. Sur C. MRRC: multiple role representation crossover interpretation for image captioning with R-CNN feature distribution composition (FDC). Multim Tools Appl. 2021;80(12):18413–43. doi:10.1007/s11042-021-10578-9.
132. Rennie SJ, Marcheret E, Mroueh Y, Ross J, Goel V. Self-critical sequence training for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017. p. 7008–24.
133. Wang W, Chen Z, Hu H. Hierarchical attention network for image captioning. Proc AAAI Conf Artif Intell. 2019;33(1):8957–64. doi:10.1609/aaai.v33i01.33018957.
134. Wu C, Yuan S, Cao H, Wei Y, Wang L. Hierarchical attention-based fusion for image caption with multi-grained rewards. IEEE Access. 2020;8:57943–51. doi:10.1109/ACCESS.2020.2981513.
135. Wu J, Chen T, Wu H, Yang Z, Luo G, Lin L. Fine-grained image captioning with global-local discriminative objective. IEEE Transact Multim. 2020;23:2413–27. doi:10.1109/TMM.2020.3011317.
136. Jiang W, Ma L, Jiang YG, Liu W, Zhang T. Recurrent fusion network for image captioning. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018. p. 499–515.
137. Thobhani A, Zou B, Kui X, Abdussalam A, Asim M, Ahmed N, et al. A concise and varied visual features-based image captioning model with visual selection. Comput Mater Contin. 2024;81(2):2873–94. doi:10.32604/cmc.2024.054841.
138. Thobhani A, Zou B, Kui X, Al-Shargabi AA, Derea Z, Abdussalam A, et al. A novel image captioning model with visual-semantic similarities and visual representations re-weighting. J King Saud Univ-Comput Inf Sci. 2024;36(7):102127. doi:10.1016/j.jksuci.2024.102127.
139. Hossen MB, Ye Z, Abdussalam A, Hossain MA. ICEAP: an advanced fine-grained image captioning network with enhanced attribute predictor. Displays. 2024;84(3):102798. doi:10.1016/j.displa.2024.102798.

140. Yang X, Tang K, Zhang H, Cai J. Auto-encoding scene graphs for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019. p. 10685–94.
141. Yao T, Pan Y, Li Y, Mei T. Exploring visual relationship for image captioning. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018. p. 684–99.
142. Xian T, Li Z, Zhang C, Ma H. Dual global enhanced transformer for image captioning. Neural Netw. 2022;148(12):129–41. doi:10.1016/j.neunet.2022.01.011.
143. Huang L, Wang W, Chen J, Wei XY. Attention on attention for image captioning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019. p. 4634–43.
144. Cornia M, Stefanini M, Baraldi L, Cucchiara R. Meshed-memory transformer for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020. p. 10578–87.
145. Luo Y, Ji J, Sun X, Cao L, Wu Y, Huang F, et al. Dual-level collaborative transformer for image captioning. Proc AAAI Conf Artif Intell. 2021;35(3):2286–93. doi:10.1609/aaai.v35i3.16328.
146. Nguyen VQ, Suganuma M, Okatani T. GRIT: faster and better image captioning transformer using dual visual features. In: European Conference on Computer Vision; 2022; Springer. p. 167–84.
147. Yu J, Wang Z, Vasudevan V, Yeung L, Seyedhosseini M, Wu Y. Coca: contrastive captioners are image-text foundation models; arXiv 2022. arXiv preprint arXiv: 220501917;2.
148. Wang Z, Yu J, Yu AW, Dai Z, Tsvetkov Y, Cao Y. SimVLM: simple visual language model pretraining with weak supervision. arXiv:210810904. 2021.
149. Li J, Li D, Savarese S, Hoi S. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: International Conference on Machine Learning; 2023; PMLR. p. 19730–42.
150. Wang J, Yang Z, Hu X, Li L, Lin K, Gan Z, et al. GIT: a generative image-to-text transformer for vision and language. arXiv:220514100. 2022.
151. Wang P, Yang A, Men R, Lin J, Bai S, Li Z, et al. Ofa: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In: International Conference on Machine Learning; 2022; PMLR; p. 23318–40.
152. Xu N, Zhang H, Liu AA, Nie W, Su Y, Nie J, et al. Multi-level policy and reward-based deep reinforcement learning framework for image captioning. IEEE Transact Multim. 2019;22(5):1372–83. doi:10.1109/TMM.2019.2941820.
153. Zhang Y, Shi X, Mi S, Yang X. Image captioning with transformer and knowledge graph. Patt Recognit Lett. 2021;143(6):43–9. doi:10.1016/j.patrec.2020.12.020.
154. Liu H, Zhang S, Lin K, Wen J, Li J, Hu X. Vocabulary-wide credit assignment for training image captioning models. IEEE Transact Image Process. 2021;30:2450–60. doi:10.1109/TIP.2021.3051476.
155. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv:181004805. 2018.
156. Zhang P, Li X, Hu X, Yang J, Zhang L, Wang L, et al. Vinvl: revisiting visual representations in vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021. p. 5579–88.
157. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I, et al. Language models are unsupervised multitask learners. OpenAI Blog. 2019;1(8):9.
158. Chen YC, Li L, Yu L, El Kholy A, Ahmed F, Gan Z, et al. Uniter: universal image-text representation learning. In: European Conference on Computer Vision; 2020; Springer. p. 104–20.
159. Hodosh M, Young P, Hockenmaier J. Framing image description as a ranking task: data, models and evaluation metrics. J Artif Intell Res. 2013;47:853–99. doi:10.1613/jair.3994.
160. Plummer BA, Wang L, Cervantes CM, Caicedo JC, Hockenmaier J, Lazebnik S. Flickr30k entities: collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE International Conference on Computer Vision; 2015. p. 2641–9.
161. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, 2014 Sep 6–12; Zurich, Switzerland: Springer. p. 13–55.
162. Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, et al. Visual genome: connecting language and vision using crowdsourced dense image annotations. Int J Comput Vis. 2017;123(1):32–73. doi:10.1007/s11263-016-0981-7.

163. du Plessis M, Brink W. Improving the performance of image captioning models trained on small datasets. In: Southern African Conference for Artificial Intelligence Research; 2021; Springer. p. 77–91.
164. Karpathy A, Fei-Fei L. Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2015. p. 3128–37.
165. Yang M, Liu J, Shen Y, Zhao Z, Chen X, Wu Q, et al. An ensemble of generation-and retrieval-based image captioning with dual generator generative adversarial network. IEEE Transact Image Process. 2020;29:9627–40. doi:10.1109/TIP.2020.3028651.
166. Yang L, Hu H, Xing S, Lu X. Constrained LSTM and residual attention for image captioning. ACM Transact Multim Comput Commun Appl. 2020;16(3):1–18. doi:10.1145/3386725.
167. Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics; 2002. p. 311–8.
168. Anderson P, Fernando B, Johnson M, Gould S. Spice: semantic propositional image caption evaluation. In: Computer Vision–ECCV 2016: 14th European Conference, 2016 Oct 11–14; Amsterdam, The Netherlands: Springer. p. 14–98.
169. Vedantam R, Lawrence Zitnick C, Parikh D. Cider: consensus-based image description evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2015. p. 4566–75.
170. Banerjee S, Lavie A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization; 2005. p. 65–72.
171. Lin CY. Rouge: a package for automatic evaluation of summaries. In: Text summarization branches out; 2004. p. 74–81.
172. Laina I, Rupprecht C, Navab N. Towards unsupervised image captioning with shared multimodal embeddings. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019. p. 7414–24.
173. Springer A, Cramer H. "Play PRBLMS" identifying and correcting less accessible content in voice interfaces. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems; 2018. p. 1–13.
174. Kalra N, Paddock SM. Driving to safety: how many miles of driving would it take to demonstrate autonomous vehicle reliability? Transport Res Part A: Pol Pract. 2016;94:182–93. doi:10.7249/RR1478.
175. Hartmann J, Holz C, Ofek E, Wilson AD. Realitycheck: blending virtual environments with situated physical reality. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems; 2019. p. 1–12.
176. Gronauer S, Diepold K. Multi-agent deep reinforcement learning: a survey. Artif Intell Rev. 2022;55(2):895–943. doi:10.1007/s10462-021-09996-w.
177. Kim T, Ahn P, Kim S, Lee S, Marsden M, Sala A, et al. NICE: CVPR 2023 challenge on zero-shot image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2024. p. 7356–65.
178. Cheng K, Song W, Ma Z, Zhu W, Zhu Z, Zhang J. Beyond generic: enhancing image captioning with real-world knowledge using vision-language pre-training model. In: Proceedings of the 31st ACM International Conference on Multimedia; 2023. p. 5038–47.
179. Ghandi T, Pourreza H, Mahyar H. Deep learning approaches on image captioning: a review. ACM Comput Surv. 2023;56(3):1–39. doi:10.1145/361759.
180. Fang Z, Wang J, Hu X, Liang L, Gan Z, Wang L, et al. Injecting semantic concepts into end-to-end image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022. p. 18009–19.