IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Deep Learning Approaches for Image Captioning: Opportunities, Challenges and Future Potential

AZHAR JAMIL [1], SAIF-UR-REHMAN [2], KHALID MAHMOOD [3], MONICA GRACIA VILLAR [4,5,6], THOMAS PROLA [4,7,8], ISABEL DE LA TORRE DIEZ [9], MD ABDUS SAMAD [10,*] AND IMRAN ASHRAF[10,*]

[1]Department of Computer Science, Barani Institute of Information Technology, PMAS-Arid Agriculture University, Rawalpindi 46000, Pakistan; (azhar@biit.edu.pk)
[2]Department of Computer Science, University Institute of Information Technology, PMAS-Arid Agriculture University, Rawalpindi 46000; (saif@uaar.edu.pk)
[3]Institute of Computing and Information Technology, Gomal University D.I.Khan, 29220, Pakistan; (khalid@gu.edu.pk)
[4]Universidad Europea del Atlántico. Isabel Torres 21, 39011 Santander, Spain; (monica.gracia@uneatlantico.es, thomas.prola@uneatlantico.es)
[5]Universidad Internacional Iberoamericana Arecibo, Puerto Rico 00613, USA.
[6]Universidade Internacional do Cuanza. Cuito, Bié, Angola
[7]Universidad Internacional Iberoamericana Campeche 24560, México.
[8]Universidad de La Romana. La Romana, República Dominicana
[9]Department of Signal Theory, Communications and Telematics Engineering. Unviersity of Valladolid, Paseo de Belén, 15. 47011 Valladolid – Spain; (isator@tel.uva.es)
[10]Department of Information and Communication Engineering, Yeungnam University, Gyeongsan 38541, Republic of Korea. (ashrafimran@live.com)

Corresponding author: Imran Ashraf and Md Abdus Samad; (Email: ashrafimran@live.com; masamad@yu.ac.kr)

This research was supported by the European University of Atlantic.

ABSTRACT Generative intelligence relies heavily on the integration of vision and language. Much of the research has focused on image captioning, which involves describing images with meaningful sentences. Typically, when generating sentences that describe the visual content, a language model and a vision encoder are commonly employed. Because of the incorporation of object areas, properties, multi-modal connections, attentive techniques, and early fusion approaches like bidirectional encoder representations from transformers (BERT), these components have experienced substantial advancements over the years. This research offers a reference to the body of literature, identifies emerging trends in an area that blends computer vision as well as natural language processing in order to maximize their complementary effects, and identifies the most significant technological improvements in architectures employed for image captioning. It also discusses various problem variants and open challenges. This comparison allows for an objective assessment of different techniques, architectures, and training strategies by identifying the most significant technical innovations, and offers valuable insights into the current landscape of image captioning research.

INDEX TERMS Image captioning; deep learning; image processing; artificial intelligence

I. INTRODUCTION

IMAGE captioning involves generating a coherent and meaningful sentence using a language model as well as visual understanding to describe the visual content of an image [1]. This task has gained significant attention recently, and the connection between natural language generation as well as human vision perception has been recently understood through neuroscience research [2]. The domain of artificial intelligence is constantly evolving and one of the recent areas of focus involves the development of architectures that can process images and generate language [3]–[5]. The primary objective of this study is to identify the most efficient channel for processing given images, representing their content, and translating them to a series of word sequences, while maintaining language fluency and establishing connections between visual and textual elements.

Initially, image captioning involved techniques such as description retrieval and hand-crafted natural language generation [6], which has been covered in previous surveys. However, with the advancements in deep learning-based generative models, image captioning has become more sophisticated [7]–[9].

It involves an image-to-sequence problem, where pixels serve as inputs. In the first step, the visual encoding stage, the inputs are transformed into one or numerous vectorized features through the process of encoding [10], [11]. The language model, in the second generating stage, uses these feature vectors to generate a string of words or sub-words from a specified vocabulary [12]. The research community has significantly improved model design in recent years. Early proposals relied on recurrent neural networks (RNN) and global image descriptors, but more recent methods include attentive approaches and reinforcement learning [13]–[15]. The latest breakthroughs include the use of transformers, self-attention, and single-stream bidirectional encoder representations from transformers (BERT)-like approaches.

Meanwhile, the computer vision and natural language processing (NLP) communities have been developing evaluation protocols and metrics to compare model results to human-generated ground truths. Despite the advancements in the field of image captioning, the task is still considered unsolved due to the lack of a single approach or solution that can accurately generate captions for images like a human. Hence, there is ongoing research to improve the quality and accuracy of the generated captions.

Different proposals and task variants have been explored to cater to various user needs and description styles in specific domains. As stated by [16] and [17], image captions can be classified into three categories: perceptual, which emphasizes visual features and non-visual attributes, which reports inferred and contextualized facts; and conceptual, which describes the real visual scene of the given content, including the relationships between visual entities.

Conceptual descriptions are typically considered the primary goal of image labeling and encompass multiple levels of aspects and details, such as including or excluding attributes or describing only targeted portions versus enhanced details. We intent to give a complete outline of the approaches, models, and task variations that have been developed in recent years to inspire novel ideas. This approach involves reviewing datasets and evaluation metrics, as well as quantitatively comparing the primary approaches. Lastly, we address the open challenges and future directions in this area.

A. RESEARCH CONTRIBUTIONS AND OBJECTIVES

This research offers a comprehensive roadmap for researchers delving into the vibrant domain of image captioning. It delivers a succinct summary of state-of-the-art techniques, diverse datasets, and robust evaluation metrics, enabling a swift recap and empowering researchers to navigate the evolving landscape of this field with informed progress. It contributes to the classifications for visual encoding and language modeling techniques, taking into account the two-fold nature of captioning models, and discusses their significant features and constraints. We examine the training methods utilized in existing research, which incorporate recent progressions achieved by pre-training and masked language model losses. Additionally, we analyze the main datasets used for image captioning, including domain-generic benchmarks and domain-specific datasets.

Further, we analyze the metrics commonly employed to evaluate performance, including both traditional and unconventional measures, and discuss the aspects of image captions that they emphasize. For this, we will perform a quantitative analysis of the primary image captioning techniques, taking into account standard as well as non-standard metrics, and analyze their relationships, performance, dissimilarities, and features. Finally, we will discuss various task variants and open challenges, as well as future directions for research.

This study presents a more extensive and current perspective of the model for caption generation using deep learning techniques in contrast to previous surveys such as [30]–[33] and [34]. We conduct a more thorough analysis of proposed approaches, examine a larger volume of literature,

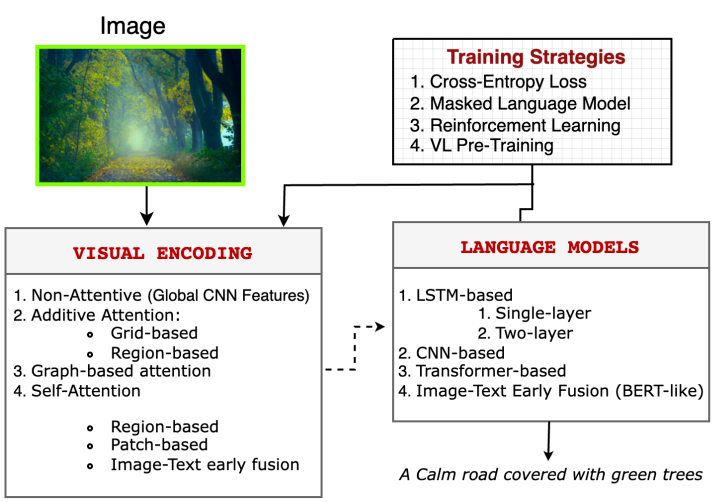


FIGURE 1: An introduction to the image-captioning task and categorization of the widely used methodologies

TABLE 1: A comparison of recent studies in image captioning with their contribution.

Ref.	Year	Contributions
[18]	2014	Introduced a deep neural network-based approach for generating captions for images using a combination of convolutional and recurrent networks.
[19]	2015	Introduced a neural network-based approach for generating captions for images using a combination of convolutional and recurrent networks.
[20]	2015	Introduced a visual attention mechanism in neural networks for image captioning to improve the quality of generated captions.
[21]	2016	Proposed a model that used convolutional neural networks to extract image features and a recurrent neural network to generate captions.
[22]	2018	Proposed a personalized image captioning model that utilized context sequence memory networks to incorporate user-specific information.
[23]	2018	Presented a bottom-up and top-down focused attention technique for captioning images that made use of both local and global picture features.
[24]	2019	Proposed a model that incorporated grounded and co-referenced people into the image captioning process to improve the quality of generated captions.
[25]	2019	Demonstrated the effectiveness of transfer learning for image captioning using a text-to-text transformer model.
[26]	2019	Proposed a model that utilized spatially and systematically attention in CNN for image-captioning.
[27]	2020	Proposed a transformer-based model for image captioning that utilized a meshed-memory architecture to handle long-term dependencies.
[28]	2021	Proposed a framework for training data-efficient image captioning models using contrastive learning and distillation techniques.
[29]	2023	A comprehensive survey summarizing and categorizing attention-based models for image captioning with the categorization of four sub-classes

and encompass unconventional evaluation criteria that are usually overlooked in other literature studies. Moreover, we consider emerging task variants and a wider range of available datasets. Figure 1 shows three broad categories of image captioning approaches discussed in this study.

A thorough comparison of numerous works in the area of picture captioning is offered in Table 1 of this research article. The table's columns for study name, contribution, and year enable quick summaries of significant research outputs. Researchers can immediately see and comprehend the improvements made in picture captioning over time thanks to the table's orderly organization of the studies and their individual contributions. It offers a detailed overview of the various methods, developments, and approaches used by various researchers as well as the development of the area.

We followed the recognized criteria for performing systematic literature reviews, including the preferred reporting items for systematic reviews and meta-analysis (PRISMA) guidelines, to ensure the comprehensiveness and rigor of our survey research. This section describes the methods we used to find relevant publications, the criteria for article selection, and the repositories we utilized to collect data.

B. LITERATURE SEARCH AND DATA COLLECTION

We searched a number of academic databases and repositories, including IEEE Xplore, ACM Digital Library, Google Scholar, Web of Science, and Scopus to find the literature relevant to image captioning. In order to uncover the most recent research findings in the area of image captioning, this search included articles published up to the year 2023.

C. INCLUSION AND EXCLUSION CRITERIA

For choosing the publications to be included in this study, we used stringent inclusion and exclusion criteria. If an article fits the following requirements, it is considered suitable for this survey

- **Relevance**: The article's main emphasis is on computer vision and natural language processing algorithms, models, or related research for image captioning.
- **Publication Date**: Articles released within the decided time range to guarantee the relevancy of the literature survey.
- **Peer-Reviewed**: In order to preserve the standards and dependability of the sources, only peer-reviewed publications are included.
- **Language**: English-language articles are taken into consideration for inclusion. The articles that do not fulfill these requirements are not included in this survey.

D. SELECTION PROCESS

The selection procedure requires several steps, including an initial screening of titles and abstracts to discover possibly relevant papers, followed by a careful examination of the complete texts of chosen articles to assess their suitability for inclusion. To reduce bias, this process is conducted separately by two or more reviewers. In order to find more pertinent publications that might not have been found during our original search, we also performed a citation analysis. This iterative process made sure that the survey contained thorough and representative research.

The paper unfolds in the following sections: Section II conducts a thorough examination of visual encoding schemes, with a specific focus on global CNN and the attention mechanism. Moving to Section III, it delineates the learning models employed for image captioning, particularly delving into LSTM-based language models. Section IV delves into the nuanced discussion of training strategies for these models. Transitioning to Section V, the spotlight is on evaluation protocols, elucidating the methodologies used. Section VI broadens the scope, presenting various Variants of Captioning. Concurrently, Section VII sheds light on the challenges and unresolved issues encountered in the field.

The paper culminates in Section VIII, encapsulating the Conclusion and paving the way for future directions.

II. VISUAL ENCODING

Visual encoding in image captioning refers to the process of transforming an input image into a compact and meaningful representation that can be easily understood by a machine learning model. It involves extracting high-level visual features from the image, which capture important information such as objects, shapes, colors, and textures [35]. The visual encoding step plays a crucial role in bridging the gap between images and natural language descriptions. Encoding the visual content of an image into a numerical representation enables the subsequent captioning model to generate accurate and relevant textual descriptions.

There are various techniques used for visual encoding, such as convolutional neural networks (CNN) and pre-trained models such as VGGNet, ResNet, or InceptionNet. These models are trained on large-scale image classification tasks and have learned to extract rich visual features from images. During visual encoding, the input image is passed through the CNN or pre-trained model, and several convolutional and pooling layers are applied to extract hierarchical visual features at different scales. The output of these layers is typically a high-dimensional map of attributes that represents the visual content of the given image.

To obtain a more compact representation, techniques such as spatial pooling or global average pooling are often applied to aggregate the spatial information across the feature map. This reduces the dimensionality of the features while preserving their semantic meaning. The resulting visual features are then fed into the subsequent captioning model, which can be a recurrent neural network (RNN) or a transformer-based architecture. The captioning model uses visual features along with a language model to generate a coherent and descriptive caption for the given image.

A. GLOBAL CNN FEATURES

In image captioning, global CNN features play a crucial role in capturing the overall content and context of an image. These features provide a compact representation of the entire image, enabling the captioning model to understand and describe the visual content effectively [36]. In this section, we will delve into the concept of global CNN features, their extraction process, and their significance in image captioning.

1) Introduction to Global CNN Features

Global CNN features are extracted from CNN and aim to capture high-level visual information that encompasses the entire image. Unlike local CNN features that focus on specific regions or objects within an image, global features consider the image as a whole and provide a holistic representation [37]. These features are derived from the output of deep convolutional layers in CNN architectures, where the layers learn to recognize complex patterns and semantic information.

2) Extraction of Global CNN Features

The process of extracting global CNN features involves passing the input image through a pre-trained CNN model, such as VGGNet, ResNet, or InceptionNet. These models are typically trained on large-scale image classification tasks, which enables them to learn rich and discriminative visual representations [38].

During the forward pass, the image undergoes a series of convolutional and pooling layers, resulting in a high-dimensional feature map. To obtain global features, spatial pooling techniques are applied, such as global average pooling or spatial pyramid pooling [39]. These techniques aggregate the spatial information across the feature map, reducing the dimensionality while retaining the essential visual content. The resulting global CNN features represent the overall visual characteristics of the image.

3) Significance of Global CNN Features in Image Captioning

Global CNN features provide a condensed representation of the image, capturing its salient visual attributes. They offer several advantages in the context of image captioning

i) **Contextual Understanding**: By considering the entire image, global features enable the captioning model to grasp the overall context and scene description. This helps in generating captions that accurately describe the visual content and provide a comprehensive understanding of the image [40].
ii) **Semantic Information**: The deep convolutional layers of CNN models are trained to recognize high-level semantic concepts, such as objects, shapes, and textures. Global CNN features encode this semantic information, allowing the captioning model to generate more descriptive and meaningful captions [41].
iii) **Robustness to Variations**: Global features are more robust to variations in object positions, scales, and occlusions within the image. They capture the most discriminative visual cues that are essential for caption generation, regardless of the specific spatial configurations of objects [42].
iv) **Efficiency**: Global features are computationally efficient compared to their local counterparts. Since they consider the entire image, the extraction process requires fewer computations and can be performed in a single forward pass through the pre-trained CNN model.

4) Integration with Captioning Models

Once the global CNN features are extracted, they serve as input to the captioning model. These features are typically combined with natural language processing components, such as RNNs or transformer-based architectures, to generate image captions [43]. The global features provide a visual context that helps the captioning model align the generated text with the visual content of the image, leading to more accurate and contextually relevant captions.

B. PAY CLOSE ATTENTION TO CNN FEATURE GRID

Many recent approaches have aimed to improve the granularity level of visual encoding [44]–[46], motivated by the constraints of global representations. Dai et al. [47] used 2D activation maps instead of 1D global feature vectors to directly include spatial structure into the language model. A significant percentage of the captioning community has embraced the additive attention method, which was inspired by machine translation literature (Figure 2b). This technique implements time-varying visual feature encoding, which provides greater flexibility and finer granularity.

A weighted averaging approach is used to characterize additive attention. In the initial formulation by Bahdanau et al. [48], attention weights are computed using a single-layer feed-forward neural network with hyperbolic tangent non-linearity. Given two vector sets $\{X_1, ..., X_n\}$ and $\{h_1, ..., h_m\}$, the additive attention score between X_i and X_j is determined as:

$$f_{att}(h_i, x_j) = W_3^T tanh W_1, x_i + W_2 x_j \qquad (1)$$

W_1 and W_2 are weight matrices, and W_3 is a weight vector that aids in linear combination. Then, using a softmax function, a probability distribution p $(x_j|h_i)$ is generated, indicating the relevance of the element represented by x_i for hh_i.

The attention mechanism, which was originally designed for sequence alignment, has been extended to connect visual representations with the hidden states of a language model. Xu et al. [45] proposed a strategy for exploiting additive attention over a convolutional layer's spatial output grid. This allows for selective focus on specific grid elements during word production. This approach has been used in other papers, with slight improvements in visual encoding [46], [49]–[52], and [53].

To enhance the encoder-decoder system, review networks have been established. Yang et al. [26] used a recurrent review network to perform many review steps with an emphasis on encoder hidden states and output a "thought vector" after each step. This vector is then used by the decoder's attention mechanism.

Chen et al. [54] advocated combining channel-wise attention with classical spatial attention over convolutional activations. They experimented with exploiting multi-level characteristics by using multiple convolutional layers. Similarly, Jiang et al. [55] proposed leveraging complementary information with several CNNs by merging their representations with a recurrent method.

Some methods incorporated human attention by combining saliency information, directing caption production, and stimulus-based attention. Sugano and Bulling [56] pioneered this concept by using human eye fixations to caption images. As an input to the soft-attention module, they provided normalized fixation histograms over the image, weighing attended visual regions based on fixation. Saliency maps were used as an additional attention source in subsequent investigations [57]–[59], and [23].

C. PAY CLOSE ATTENTION TO THE CNN FEATURE GRID

In the field of image captioning, attention mechanisms have proven to be highly effective in improving the quality and relevance of generated captions [60]. In this section, we will explore the concept of attention over a grid of CNN features, a technique that enhances the captioning process by selectively focusing on relevant visual regions within an image. This section will provide an overview of attention mechanisms, explain the grid-based approach, and discuss their significance in image captioning.

1) Introduction to Attention Mechanisms

Attention mechanisms in image captioning mimic the human visual attention process by dynamically allocating importance to different regions of an image. Rather than relying solely on a fixed global context, attention mechanisms allow the captioning model to attend to specific visual features that are most relevant to generating accurate and descriptive captions [61].

2) Grid-based Attention

The grid-based features refer to a method of applying attention mechanisms to a grid-like structure obtained from the output of the CNN model. This approach divides the feature map into a regular grid of spatial locations, each representing a specific visual region. To compute attention over the grid, a set of learnable attention weights is associated with each grid location [49]. These weights determine the importance or relevance of that region for generating the next word in the captioning process. By assigning different attention weights to different grid locations, the model can dynamically emphasize or de-emphasize specific visual regions based on their relevance. During the caption generation process, the attention mechanism attends to different grid locations at each step, allowing the model to focus on the most informative regions. This enables the captioning model to generate captions that are closely aligned with the salient visual content of the image.

3) Importance of Paying Attention to CNN Features Grid in Image Captioning

Attention to the CNN features offer several advantages in image captioning that are as follows

i) **Fine-grained Localization**: By attending to specific grid locations, the model can selectively focus on fine-grained visual details, such as objects, regions, or image-specific attributes. This fine-grained localization helps in generating more accurate and contextually relevant captions [20].

ii) **Relevance to Visual Content**: The attention mechanism enables the model to adaptively attend to relevant visual regions while generating each word in the caption. This ensures that the generated text is aligned with

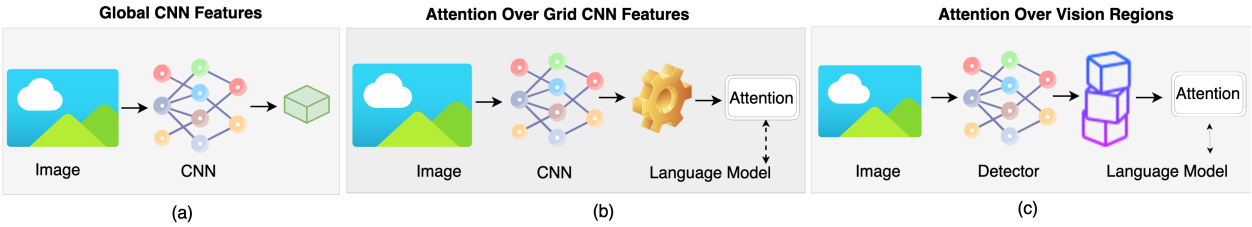


FIGURE 2: Widely used three significant visual representation techniques used for captioning of images include: (a) utilizing global CNN features, (b) employing fine-grained attributes captured from a convolutional layer, an attention mechanism supported by the language model., and (c) using an attention mechanism and image features collected by the detector.

the most important visual cues, resulting in captions that accurately describe the image content [47].

iii) **Handling Complex Scenes**: Images often contain multiple objects or complex scenes. Attention over a grid of CNN features allows the model to attend to multiple regions simultaneously, providing the ability to capture the relationships and interactions between different objects or regions within the image [52].

iv) **Interpretability**: The attention weights associated with each grid location provide insights into which regions of the image the model attends to while generating captions. This interpretability can be useful for understanding and analyzing the model's behavior and its reasoning process during caption generation [56].

4) Integration with Captioning Models

Attention over a grid of CNN features is typically incorporated into captioning models that utilize RNNs or transformer-based architectures. The attention weights obtained from the grid locations are combined with the captioning model's hidden states to generate contextually relevant and visually grounded captions.

D. ATTENTION OVER VISUAL REGIONS

Attention over visual regions refers to a mechanism used in image captioning that dynamically assigns weights or importance to different regions of an image based on their relevance to the generation of captions. This attention mechanism allows the captioning model to focus its attention on specific visual regions that are most informative and contribute significantly to the understanding and description of the image [55]. Unlike global or grid-based attention, attention over visual regions operates on a more localized level. Instead of dividing the image into a predefined grid or considering the image as a whole, attention is applied to individual regions or regions of interest (ROIs) within the image.

The process of attention to visual regions involves two main steps. First, a region proposal mechanism is used to identify and extract meaningful ROIs from the image. This can be achieved through techniques such as object detection or region-based convolutional neural networks (R-CNN).

Once the ROIs are obtained, the attention mechanism assigns attention weights to each region based on its relevance to the caption generation process. These attention weights reflect the importance of each region in contributing to the overall understanding and description of the image [23], [62], [63]. During caption generation, the captioning model selectively attends to different visual regions according to their attention weights. As a result, the model may concentrate on the most salient and informative regions while generating each word or phrase in the caption. By attending to relevant visual regions, the model can generate captions that are closely aligned with the specific content and context of the image [57]–[59]. Attention to visual regions offers several benefits in image captioning.

Relevance: By assigning attention weights to specific visual regions, the model can generate captions that are highly relevant to the content and context of the image. This ensures that the generated text accurately describes the relevant objects, scenes, or attributes within the image.

i) **Fine-grained Localization**: Attention over visual regions allows the model to focus on fine-grained details within the image, such as specific objects or regions of interest. This enables the model to generate captions that capture the intricate visual characteristics of the image [64].

ii) **Contextual Understanding**: By attending to different visual regions, the model can capture the relationships and interactions between various objects or regions within the image [65]. This enhances the model's contextual understanding and enables it to generate captions that reflect the spatial and semantic relationships present in the image.

iii) **Adaptability**: The attention mechanism allows the model to dynamically adapt its focus to different visual regions based on their importance [66]. This adaptability ensures that the model can handle images with varying complexities and generate captions that are suitable for different types of visual content.

E. GRAPH-BASED ENCODING

In recent years, graph-based encoding has emerged as a powerful technique in the field of image captioning [67], [68]. This approach leverages the inherent structural relationships

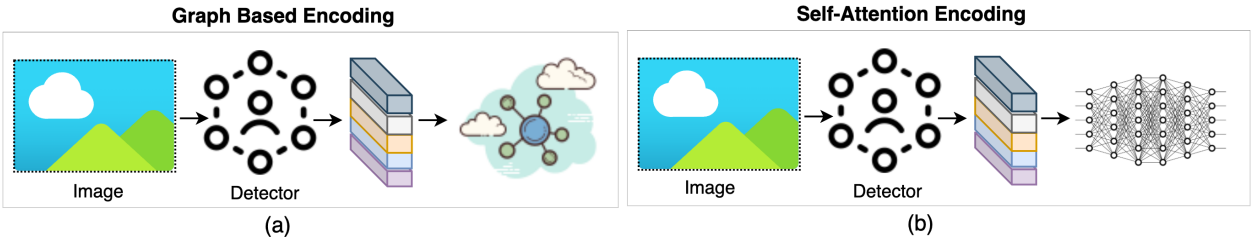


FIGURE 3: Latest advancements in visual encoding for image captioning: (a) graph-based encoding, and (b) self-attention-based encoding.

and dependencies present in visual scenes to capture rich contextual information. In this section, we will explore the concept of graph-based encoding, its methodology, and its significance in improving the quality of image captions.

Graph-based encoding in image captioning involves representing an image as a graph, where the nodes represent visual entities such as objects or regions, and the edges capture the relationships or interactions between these entities. By modeling the image as a graph, this approach allows for the explicit incorporation of spatial and semantic dependencies.

1) Methodology of Graph-based Encoding

The process of graph-based encoding typically involves several steps

i) **Object Detection**: The first step is to perform object detection or region proposal techniques to identify salient visual entities within the image. This can be achieved using pre-trained object detection models or R-CNN.

ii) **Node Representation**: Each detected object or region is represented as a node in the graph. The node representation usually consists of learned visual features extracted from the corresponding visual entity using CNN. These features encode the visual appearance, shape, and other relevant attributes of the objects or regions.

iii) **Edge Generation**: The next step is to establish edges between the nodes based on the relationships between visual entities. These relationships can be determined using heuristics, pre-defined rules, or learned from data. For example, spatial relationships such as adjacency or containment can be used to define edges between objects or regions.

iv) **Edge Representation**: Edges in the graph are associated with learned edge features that capture the contextual relationships between the connected nodes. These features can be learned from training data or derived from pre-trained models, such as graph convolutional networks (GCNs), which can capture the semantic dependencies be-tween visual entities.

v) **Graph Construction**: The nodes and edges are combined to form the complete graph representation of the image. This graph serves as a structured representation of the visual scene, encoding both the visual entities and their relationships.

2) Significance of Graph-based Encoding in Image Captioning

Graph-based encoding offers several advantages in image captioning

i) **Contextual Understanding**: By explicitly modeling the relationships between visual entities, graph-based encoding provides a more comprehensive understanding of the visual scene [69]. This enables the captioning model to generate captions that capture the contextual dependencies and interactions among objects or regions [22].

ii) **Improved Coherence**: The structured nature of the graph representation allows the captioning model to generate captions that are more coherent and globally consistent [70]. The model can leverage the graph structure to ensure that the generated captions follow logical relationships and describe the scene in a more structured and coherent manner.

iii) **Enhanced Visual Reasoning**: Graph-based encoding enables visual reasoning by explicitly encoding the semantic relationships between objects or regions [71]. This facilitates the model's ability to reason about complex visual scenes, handle ambiguous situations, and generate captions that are grounded in the underlying structure of the scene.

iv) **Adaptability to Complex Scenes**: Graph-based encoding can effectively handle complex scenes with multiple objects, occlusions, or intricate spatial arrangements. The graph structure allows the model to capture the interactions between objects, leading to more accurate and detailed captions that reflect the complexities of the visual scene.

3) Integration with Captioning Models

Once the graph representation is constructed, it is combined with captioning models such as RNN or transformer-based architectures. The graph structure is utilized as an additional input, providing the captioning model with rich contextual information.

F. SELF-ATTENTION ENCODING

Self-attention encoding, also known as self-attention mechanism or intra-attention, is a key component in many state-of-the-art natural language processing (NLP) models, including those used in image captioning [72]. It enables the model to focus on different parts of the input sequence and capture the relationships between its own elements. In this section, we explore the concept of self-attention encoding, its mathematical formulation, and its significance in enhancing the understanding and generation of image captions. Self-attention encoding is a mechanism that allows a model to compute attention weights for each element in a sequence by considering the relationships between all the elements within that sequence. It is a powerful technique for capturing dependencies and identifying important contextual information, as it enables the model to dynamically attend to different parts of the input sequence based on their relevance.

1) Mathematical Formulation

The self-attention mechanism can be mathematically expressed as follows. Given an input sequence $X = (x_1, x_2, ..., x_n)$, where x_i represents the i-th element of the sequence, we aim to compute attention weights for each element that indicates its importance or relevance to other elements in the sequence.

To perform self-attention, three types of vectors are derived from the input sequence, key vectors (Kv), a query vector (Qv), as well as value vectors (Vv). These vectors are linear projections of the input elements and are typically obtained by applying learned weight matrices.

$$Kv = XW_k, \quad Qv = XW_p, \quad Vv = XW_v \tag{2}$$

where W_k, W_p, and W_v are learnable weight matrices.

To measure the relevance between each query vector (Q) and key vector (K), attention scores are calculated. The attention score between the i-th query and j-th key is computed by taking the dot product of their respective vectors

$$Attention(Q_i, K_j) = Q_i.K_j \tag{3}$$

The attention scores undergo a normalization process using the softmax function to obtain attention weights

$$Self-Attended(X_i) = \sum Attention-Weight(Q_i, K_j).V_j \tag{4}$$

2) Significance of Self-Attention Encoding in Image Captioning

i) **Contextual Understanding**: Self-attention allows the model to capture long-range dependencies and relationships between different elements in the input sequence. This enables the model to have a better contextual understanding of the image features or textual context, resulting in more accurate and meaningful image captions.

ii) **Flexible Attention**: Unlike fixed attention mechanisms, self-attention allows the model to attend to different parts of the input sequence based on their relevance. It enables the model to dynamically adapt its focus, emphasizing important features or contextually relevant elements for generating captions.

iii) **Capture Interactions**: Self-attention captures the interactions between different elements within the sequence. In image captioning, it helps the model to capture the relationships between visual features or textual tokens, allowing for the generation of captions that are grounded in the semantic and visual coherence of the input.

iv) **Parallel Processing**: Self-attention can be computed in parallel, making it highly efficient for capturing dependencies across long sequences. This makes it suitable for processing the spatial or temporal aspects of image data, where long-range dependencies are crucial for understanding the context.

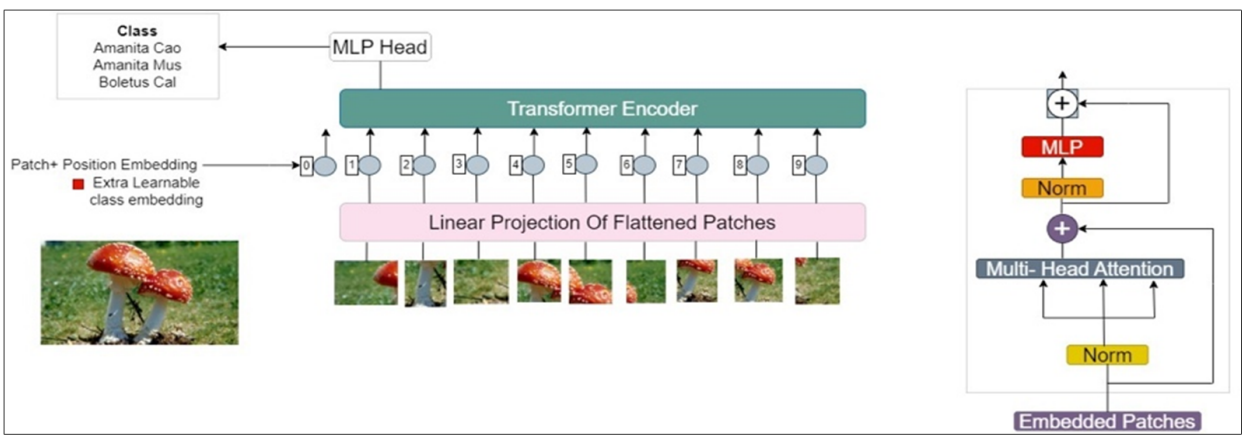


FIGURE 4: Vision T-Encoding. The picture is divided into fixed-size patches, linearly embedded, appended to position embeddings, and sent to a typical Transformer encoder.

G. DISCUSSION

The discussion on the most appropriate feature model for image captioning is evolving due to various factors. The advancements in grid features, self-attentive visual encoders, large-scale multimodal models, improved object detection, and end-to-end visual models have introduced new possibilities and challenges. The inclusion of textual information into visual representations has shown great potential and deserves further investigation. Future studies should focus on exploring and comparing these various approaches to identify the most effective feature model for image captioning tasks.

III. LEARNING MODELS FOR IMAGE CAPTIONING

Language models play a crucial role in the field of image captioning. Image captioning involves generating textual descriptions or captions that accurately convey the content and context of an image. With the advancements in deep learning and the availability of vast amounts of visual data, language models have emerged as powerful tools for enhancing the accuracy and quality of image captions

$$P(y_1, y_2, ..., y_n|X) = \Pi P(y_i|y_1, y_2, ..., y_{i-1}, X) \quad (5)$$

Equation 5 represents the probability of generating a sequence $y_1, y_2, ..., y_n$ given an input X. This equation is commonly used in the context of sequence generation tasks, such as language modeling or image captioning. $P(y_1, y_2, ..., y_n|X)$ denotes the conditional probability of generating the entire sequence $y_1, y_2, ..., y_n$ given the input X. It represents the joint probability of generating each element in the sequence, taking into account the dependencies between them and the input X. Π symbol represents the product operator, indicating that we multiply the probabilities of generating each element in the sequence together. In other words, we calculate the likelihood of generating the entire sequence by multiplying the probabilities of generating each individual element in a sequential manner.

The expression $P(y_i|y_1, y_2, ..., y_{i-1}, X)$ represents the conditional probability of generating the i-th element y_i in the sequence, given the previous elements $y_1, y_2, ..., y_{i-1}$ and the input X. It captures the dependency of each element on the preceding elements and the input, which is often modeled using techniques like RNNs or transformer models. The equation expresses the joint probability of generating a sequence by multiplying the conditional probabilities of generating each element in a sequential manner, considering the dependencies between elements and the input X. It is a fundamental formulation used in various sequence generation tasks within the realm of ma-chine learning and natural language processing.

Language models are computational models that learn and predict sequences of words or sentences based on statistical patterns in a given corpus of text. These metrics are specifically designed to capture the semantic and syntactic structures inherent in human language. By leveraging these metrics, machine learning models can generate captions that are not only coherent but also contextually relevant to the corresponding images. In the context of image captioning, language models are typically used to generate captions by encoding the visual information of an image and decoding it into a natural language representation. The language model takes the visual features extracted from the image as input and generates a sequence of words that form a descriptive caption.

There are various types of language models employed in image captioning, each with its own strengths and limitations. Some common language models used in this domain include the following.

A. RECURRENT NEURAL NETWORKS

RNNs are widely utilized for sequence modeling tasks due to their ability to capture temporal dependencies. In image captioning, RNNs play a crucial role as they process input sequences in a sequential manner, incorporating feed-back connections to retain a hidden state that preserves context from previous inputs. Among the different RNN variants, long short-term memory (LSTM) and gated recurrent unit (GRU) are widely adopted and popular choices in image captioning tasks.

1) LSTM with Single Layer

Given the sequential structure of language, RNNs naturally excel at generating sentences. Among the various RNN variants, LSTM models have emerged as the predominant choice for language modeling [73].

The architecture for captioning based on LSTM is built upon a single-layer LSTM, which was originally introduced by Vinyals et al. [19]. In this architecture, presented in Figure 5a, where the initial hidden state of the LSTM is initialized with the visual encoding. The LSTM produces the resultant caption by predicting a word at each time step. To make the prediction, the hidden state is projected onto a vector of the vocabulary size, and a SoftMax activation function is applied to obtain the probabilities of the words

i) **Additive Attention Mechanism**: The additive attention mechanism was subsequently introduced by Xu et al. [45]. As illustrated in Figure 5b, a context vector is computed by using the prior hidden state to direct the attention mechanism across the visual characteristics (X). The multilayer perceptron that forecasts the output word is then fed this context vector.

ii) **Exploring Alternative Methods**: In addition to the aforementioned single-layer LSTM architecture, several subsequent works have adopted similar decoder structures without significant architectural changes [26], [54], [74].

iii) **Visual Sentinel**: Lu et al. [46] introduced the concept of a visual sentinel, an additional learnable vector that augments the spatial image features. The decoder attends to the visual sentinel instead of the visual features when producing "non-visual" tokens such as "the,"

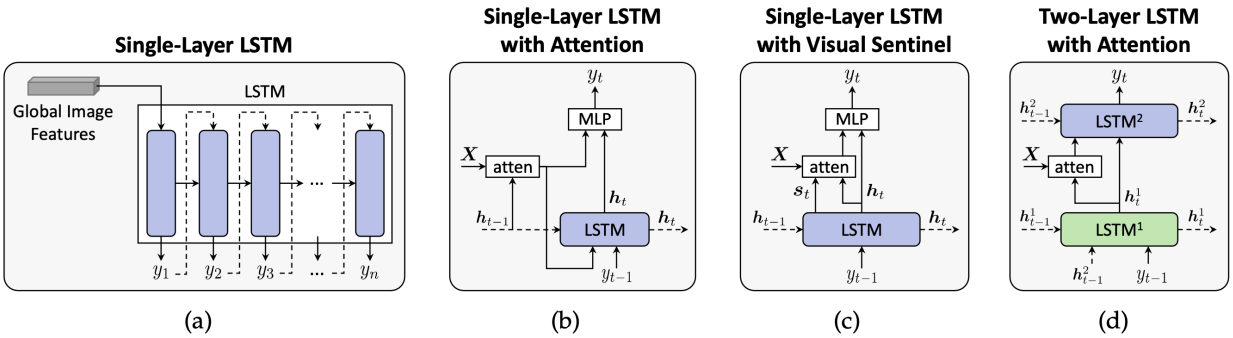


FIGURE 5: LSTM-based language modeling techniques include the following: (a) Single-Layer LSTM model conditional on the visual feature; (b) LSTM with attention, as suggested in the Show, Attend, and Tell model; (c) LSTM with attention; and (d) describe LSTM with two layers

"of," and "on" (Figure 5c). Based on the prior concealed state and the created word, the visual sentinel is calculated at each time step.

iv) **Hidden State Reconstruction**: Chen et al. [50] proposed a technique called hidden state reconstruction. This entails reconstructing the past hidden state from the present one using a second LSTM. This strategy seeks to regularize the language model's transition dynamics. The information from the two phrases of the bidirectional LSTM is combined with grid visual elements by the attention mechanism of the cross model to create the final caption.

v) **Multi-stage Generation**: Wang et al. [51] introduced the idea of multi-stage generation, where captions are generated in two different parts, generation of sentence skeleton and its attributes and are executed using one-layer LSTMs. Following a similar approach, Gu et al. [53] developed a framework coarse-to-fine multi-stage that utilizes a decoder series of LSTM. By leveraging the output of each preceding decoder, subsequent decoders in the architecture are able to refine the captions generated, resulting in a progressive improvement in the quality and accuracy of the captions.

vi) **Semantic-guided LSTM**: Jia et al. [75] suggest extension to LSTM called semantic guided LSTM. This model incorporates semantic extracted information from the image to guide the generation process. More precisely, the LSTM block incorporates semantic information as an extra input to each gate within the block.

2) Two-layer LSTM: Expanding Capabilities

To enhance their ability to capture higher-order relations, LSTM models can be extended to multi-layer structures. Donahue et al. [76] initiated a two-layer LSTM that was started to serve as a language model for captioning. Two layers are stacked in this design, and the first layer's concealed states are used as input for the second layer.

i) **Two-layers and Additive Attention**: Taking the concept further, Anderson et al. [77] advised specializing the two layers to carry out language modeling and visual attention. In Figure 5d, the top-down visual attention model is implemented in the first LSTM layer. The information from the previously created word, the prior concealed state, and the mean-pooled picture characteristics are all taken into account in this method. The second LSTM layer uses the produced attended image feature vector along with the first layer's hidden state to create a probability distribution across the vocabulary.

ii) **Alternates of Two-layer LSTM**: Due to their expressive power, two-layer LSTMs with internal attention mechanisms became widely employed as language models before the emergence of Transformer-based architectures [78]–[81].

iii) **Neural Baby Talk**: Lu et al. [82] introduced the neural baby talk approach, which incorporates a pointing network to associate words with specific image regions. The network makes predictions about certain spots or slots within the caption and inserts the appropriate picture area classes there. A visual sentinel is employed as a stand-in for non-visual words to handle for grounding reasons. The object detector is used in this method as a feature region extractor as well as a visual word prompter for the language model.

iv) **Reflective Attention**: Ke et al. [83] introduced modules for reflection into their methodology. While the second module enhances the syntactic structure of the sentence by directing the generation process based on positional information of common words, the first module computes the relevance between hidden states of all previously predicted words and the present word.

v) **Look Back and Predict Forward, Mitigating Accumulated Errors**: In a similar vein, Qin et al. [84] employed two parts of their approach. The predict ahead module concurrently predicts the following two words. By using this strategy, it is feasible to reduce the overall mistakes that may possibly occur throughout the inference process.

vi) **Time for Adaptive Attention, Dynamic Attention Steps** Huang et al. [85] devised an adaptive-attention

time mechanism that enables the decoder to take any desired number of attention steps while creating each word. By adapting the attention process in accordance with the demands of the caption creation task, this method offers flexibility.

3) Boosting LSTM with Self-Attention: Enhancing Language Models

Several studies have explored the integration of self-attention mechanisms into LSTM-based language models [27], [86]–[88], replacing the traditional additive attention mechanism. These approaches aim to improve the performance and capabilities of the models in generating captions.

Huang et al. [27] presented the operator for attention-on-attention, which augments LSTM with an additional step of attention. This operator performs attention on top of visual self-attention, allowing the model to focus on relevant visual features while generating captions.

Pan et al. [86] proposed an x-linear attention block, which incorporates interactions of second order to promote self-attention, has been proposed. This improvement makes the entire process of creating picture captions better, resulting in captions that are more accurate and contextually aware.

Presenting an alternative strategy, Zhu et al. [88] utilizes a decoder enriched with self-attention [86] to enhance the language generation process. This approach aims to automatically discover the most effective configurations for the model's architecture, resulting in improved caption quality. The inclusion of self-attention mechanisms in these studies showcases their ability to enhance the performance of LSTM-based language models in the context of image captioning. The integration of self-attention allows the models to capture more fine-grained relationships and dependencies within the visual and textual domains, leading to more accurate and contextually coherent captions.

B. TRANSFORMER MODELS

Transformer models have garnered considerable interest in the field of natural language processing. These models utilize self-attention mechanisms to effectively capture long-range dependencies within a given sequence. The transformer's encoder-decoder architecture, combined with attention mechanisms, has proven effective in generating high-quality captions.

1) Transformer-based Architectures: Revolutionizing Language Generation

This introduction of a new paradigm proposed by Vaswani et al. [89] marked a significant shift in language generation approaches. The transformer model, based on this paradigm, has become the cornerstone of various groundbreaking NLP advancements, including BERT [90] and GPT [91]. It has emerged as the de facto standard architecture for many languages understanding tasks and has also found application in image captioning.

In the case of transformer-architecture, commonly used in image captioning, the decoder plays a central role. The mechanism employs masked self-attention, with words functioning as queries and outputs from the encoder layers serving as keys and values, allowing it to attend selectively to words within the sequence. These attention mechanisms are followed by a feed-forward network Figure 6. While training, masking operation restricts the generation process to a unidirectional flow, applying attention only to previous words. Some image captioning models have employed the original transformer decoder [92]–[95]. However, researchers have put forward variations to enhance the encoding of visual features.

2) Gating Mechanisms

Li et al. [96] developed a system for gating the flow of semantic and visual information and altering representations of picture areas with semantic qualities collected from a third-party tagger. Cornia et al. [97] extended cross-attention be-

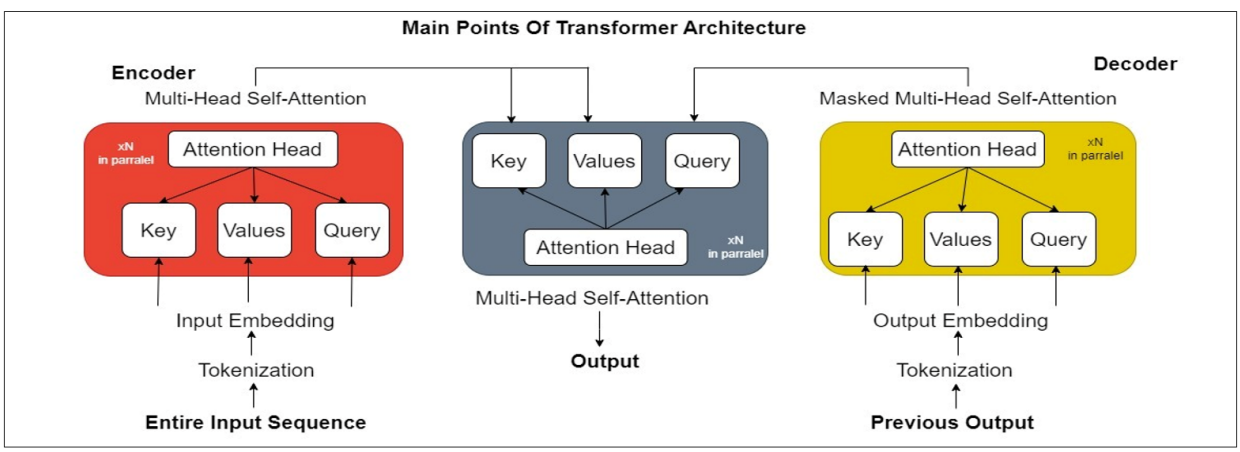


FIGURE 6: Transformer-based language model employs masked, self-attention, and cross-attention to generate captions.

yond the last encoding layer, considering all encoding layers. The transformative power of transformer-based architectures lies in their ability to capture intricate relationships between words and leverage attention mechanisms for effective language modeling. By incorporating gating mechanisms and exploring different ways of leveraging attention, researchers continue to push the boundaries of language generation and improve the integration of visual and textual information in image captioning tasks.

C. BERT-LIKE ARCHITECTURES: EXPANDING IMAGE CAPTIONING MODELS

Even though the encoder and decoder techniques are frequently employed for image captioning, recent studies have explored the integration of BERT-like structures, inspired by the BERT model [90]. These architectures fuse the visual and textual modalities early on, offering several advantages. As a result, the BERT paradigm has gained popularity in works that leverage pre-training techniques [98]–[100].

In another study, Li et al. [98] proposed the usage of tags found in images as anchor points to enhance the alignment between vision and language representations. The input image-text pair is represented by their model as a triple made up of word tokens, object tags, and region characteristics. A more reliable unified representation of vision and language is made possible by the fact that object tags match the textual classes recommended by the object detector.

By incorporating BERT-like structures, these approaches aim to leverage pre-trained models and enhance the fusion of visual and textual information in image captioning. This integration allows for more effective encoding and decoding processes, leading to improved caption quality and better alignment between the textual and the visual components.

D. NON-AUTOREGRESSIVE LANGUAGE MODELS

Non-autoregressive language models have been suggested for machine translation to shorten inference time by producing all words simultaneously. This paradigm has also been attempted to be used for the captioning of images [101]–[104]. Initially, non-autoregressive generation approaches involved multiple stages. Subsequent approaches made use of reinforcement learning techniques to enhance the final outcomes [102], [104].

E. DISCUSSION

Recurrent models have been the norm for a sizable amount of time, giving rise to creative and effective concepts that may be applied to non-recurrent solutions. Recurrent models fail to sustain long-term dependencies and have a sluggish learning curve. As a consequence of their capacity to overcome these constraints, autoregressive and transformer-based systems have become more popular. Massive pre-training for picture captioning utilizing encoder-decoder or BERT-like architectures has been developed in response to the success of pre-training on huge unsupervised corpora for NLP applications. This method has performed quite well, showing that it is possible to infer and learn visual and textual semantic relationships even from less vetted material [95], [98], [105]. While not inherently generative, BERT-like designs are ideally suited for such extensive pre-training. Examining extensive pre-training on generative-oriented architectures at the moment [95], [106] holds great promise and yields performances at least comparable to early fusion counterparts.

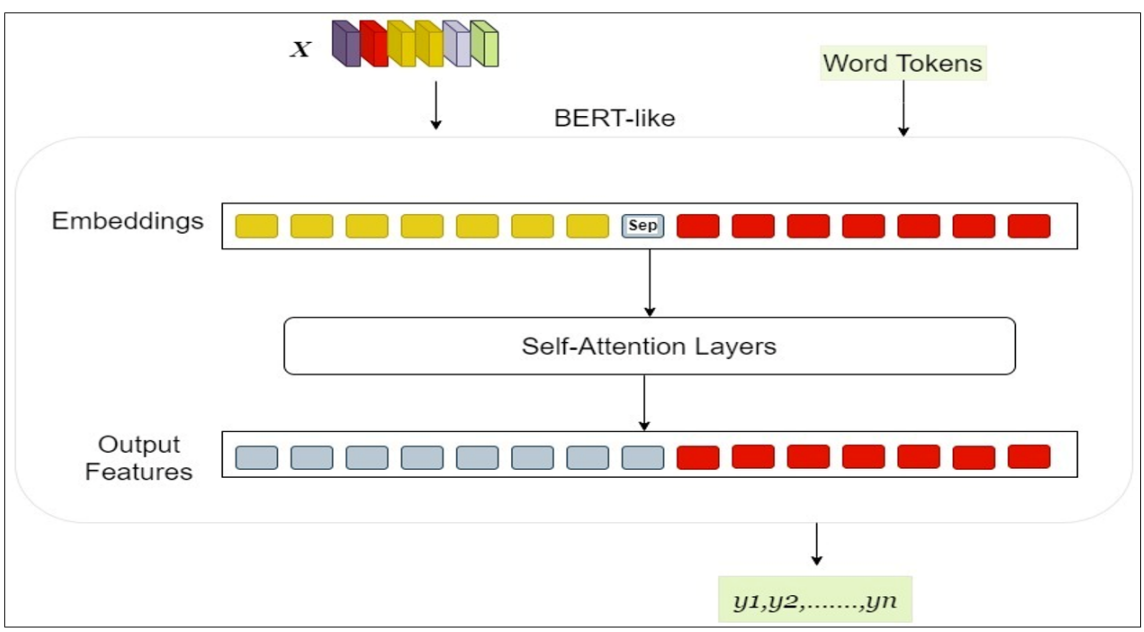


FIGURE 7: A unified stream of attentive layers in a language model that is similar to BERT simultaneously processes word tokens and image regions to produce the output caption.

IV. TRAINING STRATEGIES USED FOR IMAGE CAPTIONING MODELS

An image captioning model typically generates a caption word by word, considering both the subsequent tokens and the provided image. During each iteration, the model samples a resultant token based on the learning patterns. In this simplest process of greedy decoding, the token that has the maximum possibility is chosen as output. However, this approach accumulates prediction errors as the caption progresses. Several common training strategies are employed for this purpose

i) **Cross-entropy loss**: This strategy is based on calculating the loss using cross-entropy between the predicted word distribution and the ground truth word.
ii) **Masked language model**: This strategy involves masking some words in the input caption and training the model to predict those masked words based on the context.
iii) **Reinforcement learning**: This approach allows direct optimization, enabling the model to learn from feedback signals.
iv) **Pre-training objectives**: These objectives involve pre-training the model using vision-and-language tasks which can improve captioning performance.

A. CROSS-ENTROPY LOSS

In the context of image captioning, cross-entropy is a loss function used to train models to generate captions for images. It measures the dissimilarity between the predicted caption and the ground-truth caption. The cross-entropy loss in image captioning is calculated based on the probability distribution of predicting each word in the caption sequence. The formula for cross-entropy loss in this context is

$$LXE(\theta) = \sum log(P(y|y_i - 1, X)) \tag{6}$$

where $LXE(\theta)$ represents the cross-entropy loss, θ denotes the model's parameters, P is the probability distribution induced by the language model, y_i is the ground-truth word at time i, $y_{\{}1 \ : \ i-1\}$ represents the previous ground-truth words, and X corresponds to the visual encoding of the image.

The loss is calculated by summing the negative logarithm of the predicted probabilities for each word in the ground-truth caption sequence. The goal is to minimize this loss during training, encouraging the model to assign higher probabilities to the correct words in the caption. By optimizing the cross-entropy loss, the model learns to generate captions that closely match the ground-truth captions, as it tries to minimize the discrepancy between the predicted and actual word distributions.

B. MASKED LANGUAGE MODEL

The fundamental idea driving the optimization function is to randomly mask a small subset of tokens within the input sequence and then train the model to predict these masked tokens. This approach enables the model to utilize contextual information to infer the missing tokens, resulting in the construction of a robust sentence representation that heavily relies on contextual cues. Nevertheless, it is vital to highlight that because of its emphasis on predicting masked tokens and neglecting non-masked ones, training using this approach is relatively slower compared to training for complete left-to-right or right-to-left generation. Intriguingly, certain studies have embraced this technique as a pre-training tool, occasionally excluding cross-entropy tokens [98], [100].

C. REINFORCEMENT LEARNING IN IMAGE CAPTIONING

Researchers have turned to reinforcement learning (RL) as a viable method for training picture captioning models to get over the drawbacks of word-level training methodologies. The image captioning model is viewed as an agent in this paradigm, having movable parameters that determine its course of action. The agent applies its policy at each time step to select an action that corresponds to foreseeing the subsequent word in the created phrase. In order to maximize the predicted payoff, the agent's parameters are optimized after completion. Many research works have looked into the use of RL in picture captioning using other sequence-level metrics as reward signals. The loss gradient is computed using both beam search and greedy decoding methods as follows

$$\bigtriangledown_\theta L(\theta) = -\frac{1}{k} + \sum_{i=1}^{k}((r(w^i) - b) \bigtriangledown \theta log P(w^i)) \tag{7}$$

where w^i represents the i-th sentence in the beam or a sampled collection, $r(\cdot)$ is the reward function (e.g., CIDEr computation), and b is the baseline value.

The baseline refers to the computation of the sentence reward, which can be obtained either through greedy decoding [44] or by calculating the average reward from the beam candidates [97]. It is worth noting that RL training from a random policy is generally inefficient and time-consuming. Therefore, the typical procedure involves pre-training the model using cross-entropy or a masked language model. Subsequently, the model undergoes a fine-tuning stage with RL, incorporating a sequence-level metric as the reward. This pre-training step ensures that the initial RL policy is more favorable than a random one, leading to improved learning efficiency.

D. PRE-TRAINING IN VISION AND LANGUAGE MODELS AT A LARGE SCALE

Following the BERT strategy [90], tokens from both the visual and textual modalities are randomly masked, and the model is trained to predict the masked inputs based on the contextual information from both modalities, thereby establishing a joint representation. Another prevalent strategy involves employing a contrastive loss, where inputs are organized as triples consisting of image regions, caption words,

and object tags. The model is tasked with distinguishing between accurate triples and contaminated ones, where the tags are substituted randomly [98], [100], [114]. In some cases, cross-entropy is being utilized during training, particularly when working with inaccurate captions [95], [106], [115], [116].

The effectiveness of this technique has been shown through its ability to facilitate bi-directional attention within the prefix sequence, allowing its application in both decoder-only and encoder-decoder sequence-to-sequence models. Notably, certain large-scale models trained on inaccurate data using the given method have achieved state-of-the-art performance without the need for a reinforcement learning-based fine-tuning stage [95], [117]. Additionally, image captioning can serve as a pre-training task to efficiently learn visual representations, which in turn can provide benefits to downstream tasks like image classification, and detection of objects.

V. EVALUATION PROTOCOLS FOR IMAGE CAPTIONING

The development of image annotation depends, as with any data-driven endeavor, on the availability of huge datasets and the development of quantitative assessment criteria to evaluate performance and track improvements in the area.

A. DATASETS

Image captioning datasets comprise images paired with one or multiple captions. The inclusion of multiple ground-truth captions per image is essential for capturing the variability in human descriptions. Besides the number of available captions, the characteristics of the captions, exert a substantial influence, image captioning algorithms are profoundly shaped and their performance is greatly affected by this factor. It is crucial to consider the term distribution within the datasets since caption distributions tend to be long-tailed. Common practice involves including only terms with frequencies above a predefined threshold when using word-level dictionaries. However, subword-based tokenization approaches like BPE [118] have gained popularity, as they allow dataset pre-processing to be avoided. The datasets available vary in terms of the images they contain and the labels linked to those images. Table 2 gives a reflection of the most widely used public datasets.

Please note that some information, such as the vocabulary size and the number of words per caption, may not be available for certain datasets. Additionally, newer datasets or updates to existing datasets might also have been released. Therefore, it is always a good idea to consult the latest research papers and resources for the most up-to-date information on image captioning datasets.

1) Commonly Used Datasets for Standard Image Captioning Tasks

Researchers in the community employ benchmark datasets to enable comparisons between various approaches on a shared evaluation platform. This methodology aids in guiding the advancement of image captioning techniques by identifying appropriate directions. Benchmark datasets must accurately represent the task, encompassing its challenges and the desired optimal outcomes, which align with performance levels attainable by humans. Moreover, these datasets should include a large variety of images from different domains, and each image should have multiple associated captions.

Early structures for image captioning [18], [76], [119] typically underwent training and evaluation using the Flickr30K [108] and Flickr8K [16] datasets. These datasets encompass images sourced from the Flickr website, portraying everyday activities, events, and scenes, along with five captions per image. Presently, the most extensively utilized dataset is Microsoft COCO [120], comprising images featuring complex scenes with people, animals, and common everyday objects within their contextual environment. It contains more than 121,000 photos, each of which has five precisely written subtitles. The dataset is split into 40,504 validation pictures and 82,583 training images. For evaluation purposes, most research literature follows the splits defined by Karpathy et al. [18], wherein 5,500 images from the original validation set are assigned for validation, 5,500 for testing, and the remainder for training. The dataset comprises 40,700 images, with each image having 45 private captions, and it is accompanied by a public evaluation server.

2) Pre-training Datasets

Although using big, carefully curated datasets for training is a reasonable strategy, research works [95], [98], [105], [121] show the advantages of pre-training on increasingly bigger

TABLE 2: Overview of image captioning datasets.

Dataset	Domain	No. of Images	No. Of Caps. (per image)	Vocab Size	No. Of Words (per caps)	Year
Conceptual Captions [107]	Generic	3.3M	5	28K	10.3	2018
MS COCO [69]	Generic	123K	5	10K	6.2	2014
Flickr30k [108]	Generic	31K	5	-	6.1	2014
SBU Captions [4]	Generic	1K	5	-	9.1	2011
Visual Genome [64]	Generic	108K	-	-	-	2017
VizWiz Captions [109]	Assistive	31K	5	-	11.9	2018
CUB-200 [110]	Birds	11K	10	-	-	2011
Oxford-102 [111]	Flowers	8K	10	-	-	2014
Fashion Captions [112]	Fashion	52K	5	-	-	2019
BreakingNews [113]	News	1.2K	5	-	-	2018

vision and language datasets. These datasets may be gathered for different purposes, such as visual question answering, or they may be picture captioning datasets with lower-quality captions [98], [99] text-to-image generation [122], or image-caption association [123]. Prominent datasets specifically curated for image captioning pre-training purposes comprise SBU Captions, initially employed for image captioning in a retrieval context, encompassing roughly 1 million image-text pairs scraped from the Flickr website. Another dataset, YFCC100M [124], comprises 100 million media objects, with around 14.8 million images alongside automatically gathered textual descriptions. Conceptual Captions [107], [125] datasets have been proposed, offering about 3.4 million (CC3M) and 12 million (CC12M) images coupled with weakly-associated descriptions automatically collected from the web. Furthermore, the Wikipedia-based Image Text (WIT) [126] dataset offers images from Wikipedia alongside various extracted metadata from the original pages, featuring approximately 5.3 million images with corresponding English descriptions. These datasets are particularly intriguing for pre-training due to their large scale and diverse caption styles. However, it is worth noting that the captions in these datasets may contain noise. Additionally, as many photos are offered as URLs, their availability is not always guaranteed. Pre-training on these datasets requires a significant investment in processing power and time to gather the necessary data. Nevertheless, this approach proves valuable in achieving state-of-the-art performance. Certain pre-training datasets, such as ALIGN [95], [127], and ALT-200 [105], 1.8 billion and 200 million noisy image-text pairings, respectively, are not available to the general public. Moreover, the datasets used for training DALL-E [122] and CLIP [123], comprising 251 million and 401 million pairs, respectively, are also not publicly available.

3) Datasets of Specific Domains

While benchmark datasets that are generic to various domains capture the fundamental aspects of image captioning, domain-specific datasets play a critical role in illuminating and addressing specific challenges. These challenges might revolve around the visual domain, such as image type and style, or the semantic domain. The distribution of terms used to describe domain-specific images can significantly differ from that of terms used for generic images. An instance of a domain-specific dataset within the visual domain is the VizWiz Captions [109] dataset, designed to advance image captioning research in assistive technologies. This dataset comprises images taken by visually-impaired individuals using their phones, which might result in low-quality images capturing a wide range of everyday activities, many of which involve reading text. Examples of datasets in specific semantic domains include CUB-200 [110] and Oxford-102 [111]. The CUB-200 dataset contains images of birds, while the Oxford-102 dataset features images of flowers. Both datasets include ten captions per image, curated by Reed et al. [128]. Due to their specificity, these datasets are often utilized for tasks beyond standard image captioning, like cross-domain labeling [129]–[133].

Fashion Captioning [112] is another domain-specific dataset that comprises images of clothing items in different poses and colors, sometimes sharing the same caption. The vocabulary used to describe these images is typically smaller and more specific than that used in generic datasets. Conversely, datasets like Breaking News [113] and Good News [134] require enriched vocabulary as their images, extracted articles from news, and feature long labels written by multiple journalists. TextCaps [135], a dataset containing images with text that must be "read" and incorporated into the caption, and Localized Narratives [136], featuring captions narrated by people describing what they see in the images, are additional examples of domain-specific datasets. The collection of domain-specific datasets and the development of solutions to tackle the challenges they present are crucial for expanding the applicability of image captioning algorithms.

TABLE 3: An examination of the performance of representative image captioning methods concerning various evaluation metrics. The † marker denotes models trained by us using ResNet-152 features, while the ‡ marker signifies unofficial implementations.

Model	# Params (M)	B-1	B-4	M	R	C	S	Div. 1	Div. 2	Vocab	Novel	WMD	Alignment	Coverage	TIGEr	BERT-S	CLIP-S	CLIP-SRef
Show and Tell† [19]	13.6	72.4	31.4	25.0	53.1	97.2	18.1	0.014	0.045	635	36.1	16.5	0.199	71.7	71.8	93.4	0.697	0.762	
SCST (FC)‡ [44]	13.4	74.7	31.7	25.2	54.0	104.5	18.4	0.008	0.023	376	60.7	16.8	0.218	74.7	71.9	89.0	0.691	0.758	
Show, Attend † [45]	18.1	74.1	33.4	26.2	54.6	104.6	19.3	0.017	0.060	771	47.0	17.6	0.209	72.1	73.2	93.6	0.710	0.773	
SCST (Att2in)‡ [44]	14.5	78.0	35.3	27.1	56.7	117.4	20.5	0.010	0.031	445	64.9	18.5	0.238	76.0	73.9	88.9	0.712	0.779	
Up-Down‡ [77]	52.1	79.4	36.7	27.9	57.6	122.7	21.5	0.012	0.044	577	67.6	19.1	0.248	76.7	74.6	88.8	0.723	0.787	
SGAE [79]	125.7	81.0	39.0	28.4	58.9	129.1	22.2	0.014	0.054	647	71.4	20.0	0.255	76.9	74.6	94.1	0.734	0.796	
MT [80]	63.2	80.8	38.9	28.8	58.7	129.6	22.3	0.011	0.048	530	70.4	20.2	0.253	77.0	74.8	88.8	0.726	0.791	
AoANet [27]	87.4	80.2	38.9	29.2	58.8	129.8	22.4	0.016	0.062	740	69.3	20.0	0.254	77.3	75.1	94.3	0.737	0.797	
X-LAN [86]	75.2	80.8	39.5	29.5	59.2	132.0	23.4	0.018	0.078	858	73.9	20.6	0.261	77.9	75.4	94.3	0.746	0.803	
DPA [87]	111.8	80.3	40.5	29.6	59.2	133.4	23.3	0.019	0.079	937	65.9	20.5	0.261	77.3	75.0	94.3	0.738	0.802	
AutoCaption [88]	-	81.5	40.2	29.9	59.5	135.8	23.8	0.022	0.096	1064	75.8	20.9	0.262	77.7	75.4	94.3	0.752	0.808	
ORT [92]	54.9	80.5	38.6	28.7	58.4	128.3	22.6	0.021	0.072	1002	73.8	19.8	0.255	76.9	75.1	94.1	0.736	0.796	
CPTR [117]	138.5	81.7	40.0	29.1	59.4	129.4	-	0.014	0.068	667	75.6	20.2	0.261	77.0	74.8	94.3	0.745	0.802	
M^2 Transformer [97]	38.4	80.8	39.1	29.2	58.6	131.2	22.6	0.017	0.079	847	78.9	20.3	0.256	76.0	75.3	93.7	0.734	0.792	
X-Transformer [86]	137.5	80.9	39.7	29.5	59.1	132.8	23.4	0.018	0.081	878	74.3	20.6	0.257	77.7	75.5	94.3	0.747	0.803	
Unified VLP [99]	138.2	80.9	39.5	29.3	59.6	129.3	23.2	0.019	0.081	898	74.1	26.6	0.258	77.1	75.1	94.4	0.750	0.807	
VinVL [100]	369.6	82.0	41.0	31.1	60.9	140.9	25.2	0.023	0.099	1125	77.9	20.5	0.265	79.6	75.7	88.5	0.766	0.820	

B. EVALUATION METRICS

Assessing the quality of generated captions poses a challenge as it is subjective and complex. Captions must not only be grammatically correct and fluent but also accurately describe the input image. During the process of planning a human evaluation campaign involving multiple users scoring the produced sentences is widely considered one of the most reliable methods for evaluating caption quality, it is costly and lacks reproducibility, limiting fair comparisons between different approaches. Automatic scoring methods have been developed to check the quality of produced captions by the system. Table 3 provides an overview of the evaluation metrics. It presents the results of various image captioning approaches, analyzing their performance using different scores for evaluation as discussed previously. In addition, we present the parameter count to offer an understanding of the computational complexity and memory usage of the models. The information in Table 3 is derived from the model weights and captions files provided by the original authors or other top-performing implementations. The evaluation is performed on the domain-generic COCO dataset, widely used as a benchmark in the field.

Table 3 clusters the methods based on visual encoding information and orders them by CIDEr score. We see that the addition of region-based visual encodings considerably enhanced conventional and embedding-based measures. More enhancement came from integrating information on inter-object relations expressed through graphs or self-attention. CIDEr, SPICE, and Coverage metrics demonstrate the most notable benefits of vision and language pre-training. Additionally, diversity-based scores, such as Div-1 and Div-2, show a strong correlation with Vocab Size. The learning-based scores show that models trained exclusively on textual input do not successfully differentiate across picture captioning techniques. This property is desirable for image captioning evaluation, enabling performance estimation without relying on limited and subjective reference captions.

1) Benchmark Evaluation Metrics

Initially, image captioning performance was evaluated using NLP tasks like, the BLEU score [137] and METEOR score [138] were introduced in machine-translation evaluation. BLEU calculates n-gram precision up to length four, while METEOR prioritizes the recall of matching unigrams found in the candidate captions and reference sentences, considering stemming. The ROUGE score [139] has also been used for image captioning. This score considers the longest subsequences of tokens in the same relative order that appear in both the candidate and reference captions, possibly with other tokens in between. Later on, dedicated metrics for evaluating image captioning were introduced [140], [141]. The CIDEr score [141] quantifies the cosine similarity between the term frequency-inverse document frequency (TF-IDF) weighted n-grams found in the candidate caption and the set of reference captions linked to the image. It considers both precision and recall. On the other hand, the SPICE score [141] assesses matching tuples extracted from the candidate caption and the reference (or image) scene graphs, giving priority to semantic content over fluency. Metrics explicitly tailored for image captioning exhibit better alignment with human judgment when compared to metrics borrowed from other NLP tasks (with the exception of METEOR [138]), both on the corpus and caption levels [138], [141]. The correlation with human judgment is gauged through statistical correlation coefficients like Pearson's, Kendall's, and Spearman's, along with agreement with humans' preferred captions in pairs of candidates, assessed on a selected set of captioned images.

2) Metrics for assessing diversity

In order to evaluate performance, it is customary to consider a set of standard metrics as mentioned above. However, these metrics can be manipulated as they prioritize word similarity rather than the correctness of meaning. Another limitation of standard metrics is their inability to capture and encourage the generation of novel and diverse captions, which aligns better with the variability seen in human descriptions of complex images. To address this concern, diversity metrics have been developed. These metrics, such as [142]–[145], can perhaps be determined even without ground-truth captions during testing. It is advised to combine them with other metrics because they do not account for the syntactic correctness or relevance of the captions to the image. When numerous captions are produced for the same picture, a captioning system's overall effectiveness can be measured in terms of corpus-level diversity or single-image diversity (referred to as global diversity and local diversity, respectively, in [143]).

3) Embedding-based Metrics

A different approach to evaluate image captions entails the use of metrics based on embedding [146]–[149] which assess the semantic similarity or specific aspects of caption quality. This approach considers the use of embeddings to measure the degree of similarity or evaluate particular elements within the generated captions. For example, the WMD score [150], initially developed to assess semantic dissimilarity between documents, can be modified for captioning evaluation by comparing generated captions with ground-truth captions as the documents being compared [151]. In a similar vein, the Alignment score [152] analyzes whether ideas are mentioned in a human-like order by comparing the alignment of noun sequences in candidate and reference sentences. Moreover, the Coverage score [153], [154] calculates the extent of a caption by taking into account the scene's indicated visual elements. This score directly considers visual elements and can be used even without ground-truth captions.

4) Evaluation Through Learning-based Methods

As a further advancement in caption quality assessment, researchers are exploring learning-based evaluation strategies [155]–[160]. This approach employs a component within a complete captioning system responsible for assessing the

completeness [161] or human-likeness [162] of the generated captions. As an alternative, learning-based assessment uses a model that has already been trained. Consider the BERT-S score [163], which is frequently used to assess different language-generating tasks [164], uses cosine similarity to represent and contrast the tokens in the reference and candidate sentences using pre-trained BERT embeddings [90]. Additionally, the CLIP-S score [165] computes an adjusted cosine similarity between the image and candidate caption representations to evaluate image captioning directly using the CLIP [123] model. As a result, CLIP-S can work without reference captions, however, the CLIP-SRef variation can also use them.

VI. VARIANTS OF IMAGE CAPTIONING

In addition to general-purpose image captioning, the literature has explored several specific sub-tasks that can be categorized into four distinct categories based on their focus.

A. ADDRESSING SCARCITY OF TRAINING DATA

Obtaining datasets containing pairs of images and captions can incur significant expenses. Consequently, researchers have explored variants of image captioning that require less supervision.

1) Novel Object Captioning

This alternative seeks to characterize objects that do not exist in the training set, facilitating zero-shot learning to enhance real-world practicality [166], [167]. In earlier attempts [168], [169], knowledge transfer from out-domain images was explored by incorporating external unpaired visual and textual data during model training. To facilitate further advancements in this domain, a more demanding dataset called nocaps [170] was introduced, which includes nearly 400 novel objects. Newer methodologies for this specific variation [171], [172] incorporating coping mechanisms into the language model have been part of the approach.

By conditioning the model on external, unpaired visual and textual input during training, early methods tried to transfer information from out-of-domain pictures. The more difficult Nocaps dataset has been made available to aid study in this area [170], and has been released, comprising almost 401 unique items. Some methods for this type [171], [172] include copying mechanisms in the language model to choose novel things that have been anticipated by a tagger or to create a caption template with placeholders for novel objects. Additionally, based on the predictions of a tagger, Anderson et al. [170] created the Constrained Beam Search method, which guarantees the presence of chosen tag words in the output caption. In addition, Hu et al. [173] introduced a multi-layer transformer model that has been pre-trained by randomly masking one or more tags from image-tag pairings, continuing the pre-training trend with BERT-like architectures.

2) Captioning for Unpaired Images

Unsupervised and semi-supervised approaches may both be used for unpaired image captioning. Without using paired image-text training data, unsupervised captioning focuses on understanding and characterizing pictures. Early investigations and unpaired machine translation algorithms were used as models [174] suggests generating captions in a pivot language and then translating them to the target language.

Subsequently, after that, the emphasis turned to adversarial learning by teaching an LSTM-based discriminator to tell the difference between genuine captions and produced captions [175], [176]. The creation of captions from picture scene-graphs [177] and the use of memory-based networks [178] are other methods. Adversarial learning is used in semi-supervised methods like [179] that use both paired and unpaired data, whereas iterative self-learning is used in [180].

3) Continual Captioning

Following the ideas of the constant learning paradigm, continuous captioning tries to overcome the problem of handling partially unavailable material by enabling progressive learning of new tasks without losing previously acquired ones. This method views new assignments as collections of captioning tasks, each requiring the use of a different vocabulary [181], requiring the model to transfer visual concepts between tasks.

B. CONSIDERING THE VISUAL INPUT

Certain subtasks focus on establishing a stronger correlation between textual descriptions and visual data.

1) Dense Captioning

Johnson et al. [182] introduced the concept of dense captioning, which entails simultaneously identifying and defining salient regions in the image using brief sentences of natural language. This task can be compared to an expansion of object detection, where captions replace object tags, or image captioning, where specific regions replace the entire image. Approaches that take on this task make use of attribute generators [183], [184], contextual and global features [185], [186], and textual paragraphs to provide a coherent story about the visual content [187]–[192].

2) Text-based Image Captioning

Text-based image captioning, also known as optical character recognition (OCR)-based image captioning or image captioning with reading comprehension, involves reading and incorporating text present in images into the generated descriptions. Sidorov et al. [135] introduced this task with the TextCaps dataset, while OCR-CC [193], a sub-part of the CC3M dataset [107], was designed for pre-training purposes. The typical method for this variation entails integrating image regions and text tokens, leveraging their mutual spatial information [194], [195], within the visual labeling [135], [196], [197].

3) Change Captioning

Change captioning is centered around capturing alterations that transpire in a scene, necessitating precise change detection and effective natural language depiction. The Spot-the-Diff dataset [198], was introduced to tackle this task, comprising pairs of frames from video surveillance footage, accompanied by corresponding textual descriptions of visual changes. To delve deeper into this variation, the CLEVR-Change dataset was developed [199], containing nearly 80K image pairs with five different scene change types. Proposed approaches for change captioning apply attention mechanisms to emphasize semantically relevant aspects Ignoring distracting factors, such as alterations in viewpoint [200]–[202]. Additionally, some approaches are discussed in [203], where the objective is to retrieve an image based on its corresponding image and the description of the changes that took place. These variants of image captioning explore different aspects such as dealing with limited training data, enhancing the correlation between textual and visual information, and capturing changes within a scene.

C. FOCUSING ON THE TEXTUAL OUTPUT

Every image captures a wide range of entities with intricate interactions, leading to diverse human descriptions that are grounded in different objects and details. Some image captioning variants specifically target these aspects.

Diverse image captioning seeks to mimic the level of detail and variety in phrases created by people. The most popular method for achieving variety is based on several beam search algorithms [204]. This approach involves dividing the beams into similar groups and encouraging diversity between these groups. Other solutions, such as contrastive learning [205], conditional GANs [142], [162], and paraphrasing [206], have also been explored. Nevertheless, these approaches frequently exhibit subpar performance regarding caption quality, a drawback that can be partly mitigated by employing variational autoencoders [207]–[210].

An alternative approach entails using multiple part-of-speech tag sequences, predicted from image region classes [211], to prompt the model into generating diverse captions based on these sequences. While some approaches focused on cross-lingual image captioning [212], [213], captioning for medical images [214]–[216], paintings [217], [218], and news [219]–[221] has also been explored. Similarly, personalized image captioning for social networks has also been investigated [22], [222]–[225].

D. HANDLING REQUIREMENTS OF USERS

When captions do not state the obvious and are written in an interesting way that piques users' attention, users find them to be more effective. This need is addressed by personalized picture captioning, which creates descriptions based on the user's past knowledge, active vocabulary, and writing style. Earlier techniques used a memory block to retain context [226], [227]. Another approach proposed by Zhang et al. [228] used the multi-modal approach. By taking into account the user's most recent captions and a trained user representation, a transformer network is used to customize captions. Some research focuses on adding style to captions as an additional controlled input, using corpora of unpaired stylized text [229]–[233].

1) Controllable Captioning

Controllable captioning involves active user involvement, where users select and prioritize images that should be depicted with relevant details, which function as a guiding signal during the caption generation process. The guiding signal in image captioning may have sparsity, appearing as chosen image regions (as seen in [152], [234]) or words provided by users [211].

Alternatively, it can also manifest as a dense signal, such as mouse traces [136], [235]. Additionally, it can integrate some structure, like sequences that encode the order of mentioned concepts (part-of-speech tags, as in [211]) or visual objects [152]. The guiding inputs can encapsulate the user's interest in object relationships, as exemplified through the use of verbs and semantic roles to depict activities within the image and the involvement of objects in these activities [236]–[239]. Alternatively, using scene graphs generated or selected by users can also be used [237], [238].

2) Editing of Image Captioning

Sammani et al.'s [240] introduction of image captioning modification recognizes the existence of repeats and discrepancies in produced captions. In this method, the separation of the decoding stage into the two separate processes of caption production—caption generation and caption polishing is emphasized. In the last phase, the created captions are fixed of syntactic mistakes [241].

VII. CHALLENGES AND UNRESOLVED ISSUES

Despite tremendous advances, image captioning faces a number of hurdles and unresolved issues, necessitating ongoing research to enhance the subject. These problems highlight places where present techniques fall short and highlight the need for innovation to improve the sophistication of image captioning algorithms.

One of the most difficult challenges is to capture temporal correlations in image sequences or films in order to provide coherent and temporally appropriate captions. It is necessary to improve models in order to effectively incorporate temporal linkages for dynamic scenes and occurrences. It is critical for proper captioning to develop ways to account for time-dependent contextual changes.

A key barrier is the semantic gap between visual content and textual descriptions. It is still difficult to extract sophisticated semantic features from photographs and present them coherently in spoken language. Bridging the semantic gap to generate contextually rich and nuanced captions remains a challenge. The alignment of visual and verbal representations is being refined through research.

Images having ambiguous content, which can have various interpretations, provide difficulties for image captioning models. It is an ongoing concern to resolve ambiguity in descriptions and guide models to adopt the most contextually suitable interpretation. More research is needed to improve models' ability to handle confusing conditions. Captions for images frequently lack contextual information, resulting in descriptions that may not adequately represent relationships between visual components. Improving contextual comprehension in captioning models to account for complex situations, temporal linkages, and contextual signals is a research goal that will continue to be pursued. Integrating temporal data is critical for complete contextualization.

Because of the subjective nature of language and the lack of commonly established measures, evaluating the quality of generated captions is difficult. Creating solid evaluation measures that are aligned with human judgment and capable of capturing the diversity of appropriate captions is still a work in progress. Progress in evaluation methodology is critical for accurate progress assessment. The need for quick processing makes real-time image captioning difficult, especially in dynamic or live scenarios. For practical applications, balancing speed and precision is critical.

Addressing these issues will necessitate a collaborative and interdisciplinary effort. Continued exploration of novel techniques, advancements in model architectures, and a better understanding of the intricate relationship between visual and linguistic elements, including temporal dynamics, are required for image captioning to progress toward greater accuracy, diversity, and contextual relevance.

VIII. CONCLUSION AND FUTURE DIRECTIONS

This research has offered a thorough examination of the multidimensional environment of picture captioning, including a thorough examination of methods, datasets, and evaluation metrics. This complex tapestry of research demonstrates the ongoing evolution of strategies and approaches for bridging the semantic gap between visual material and natural language descriptions.

We examined a wide range of image captioning methods, from early approaches based on hand-crafted features to the most recent advances in deep learning-based generative models. The use of attention processes, multi-modal connections, and transformer architectures has greatly increased the ability to generate coherent and contextually appropriate captions. This survey serves as a road map for researchers and practitioners, providing insights regarding methodology evolution and strengths.

A thorough evaluation of datasets found that they play a critical role in building and benchmarking picture captioning models. The survey has highlighted the diversity of data sources accessible, from pioneering datasets to modern domain-specific benchmarks. As picture captioning models progress, the need for diverse and representative datasets becomes more obvious. The curation of datasets that capture the complexity and nuances of real-world settings should be the focus of future research. The survey examined the various evaluation measures used to assess the performance of captioning models. This paper attempts to aid researchers in selecting acceptable evaluation procedures by considering both established metrics and developing alternatives. Standardization and consensus on evaluation metrics are imperative for fostering fair comparisons between different approaches. Future endeavors should focus on refining existing metrics and potentially introducing novel measures that better capture the nuances of caption quality.

The future direction involves investigations into temporal relationships, real-time captioning, and cross-modal representations, all of which offer intriguing avenues for developing more dynamic and responsive captioning systems. Collaborative efforts to standardize evaluation methods and curate varied datasets will act as guiding lights for objective benchmarking, supporting continual progress in this dynamic sector. Putting an emphasis on multidisciplinary approaches and creative solutions will surely be critical in propelling picture captioning to greater practical relevance and success in real-world applications.

REFERENCES

[1] J.-Y. Pan, H.-J. Yang, P. Duygulu, and C. Faloutsos, "Automatic image captioning," in 2004 IEEE International Conference on Multimedia and Expo (ICME)(IEEE Cat. No. 04TH8763), vol. 3. IEEE, 2004, pp. 1987–1990.

[2] A. Ardila, B. Bernal, M. Rosselli et al., "Language and visual perception associations: meta-analytic connectivity modeling of brodmann area 37," Behavioural neurology, vol. 2015, 2015.

[3] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, "Every picture tells a story: Generating sentences from images," in Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11. Springer, 2010, pp. 15–29.

[4] V. Ordonez, G. Kulkarni, and T. Berg, "Im2text: Describing images using 1 million captioned photographs," Advances in neural information processing systems, vol. 24, 2011.

[5] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," Advances in neural information processing systems, vol. 26, 2013.

[6] A. Gupta, Y. Verma, and C. Jawahar, "Choosing linguistics over vision to describe images," in Proceedings of the AAAI conference on artificial intelligence, vol. 26, no. 1, 2012, pp. 606–612.

[7] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," arXiv preprint arXiv:1411.2539, 2014.

[8] A. Karpathy, A. Joulin, and L. F. Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping," Advances in neural information processing systems, vol. 27, 2014.

[9] B. Z. Yao, X. Yang, L. Lin, M. W. Lee, and S.-C. Zhu, "I2t: Image parsing to text description," Proceedings of the IEEE, vol. 98, no. 8, pp. 1485–1508, 2010.

[10] A. Aker and R. Gaizauskas, "Generating image descriptions using dependency relational patterns," in Proceedings of the 48th annual meeting of the association for computational linguistics, 2010, pp. 1250–1258.

[11] Y. Yang, C. Teo, H. Daumé III, and Y. Aloimonos, "Corpus-guided sentence generation of natural images," in Proceedings of the 2011 conference on empirical methods in natural language processing, 2011, pp. 444–454.

[12] S. Li, G. Kulkarni, T. Berg, A. Berg, and Y. Choi, "Composing simple image descriptions using web-scale n-grams," in Proceedings of the fifteenth conference on computational natural language learning, 2011, pp. 220–228.

[13] M. Mitchell, J. Dodge, A. Goyal, K. Yamaguchi, K. Stratos, X. Han, A. Mensch, A. Berg, T. Berg, and H. Daumé III, "Midge: Generating

image descriptions from computer vision detections," in Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, 2012, pp. 747–756.

[14] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, "Babytalk: Understanding and generating simple image descriptions," IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 12, pp. 2891–2903, 2013.

[15] P. Kuznetsova, V. Ordonez, T. L. Berg, and Y. Choi, "Treetalk: Composition and compression of trees for image descriptions," Transactions of the Association for Computational Linguistics, vol. 2, pp. 351–362, 2014.

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

[17] N. Sharif, U. Nadeem, S. A. A. Shah, M. Bennamoun, and W. Liu, "Vision to language: Methods, metrics and datasets," Machine Learning Paradigms: Advances in Deep Learning-based Technological Applications, pp. 9–62, 2020.

[18] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3128–3137.

[19] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3156–3164.

[20] Z. Gan, C. Gan, X. He, Y. Pu, K. Tran, J. Gao, L. Carin, and L. Deng, "Semantic compositional networks for visual captioning," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 5630–5639.

[21] M. Najman, "Image captioning with convolutional neural networks," 2017.

[22] C. Chunseong Park, B. Kim, and G. Kim, "Attend to you: Personalized image captioning with context sequence memory networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 895–903.

[23] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "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, pp. 6077–6086.

[24] A. Rohrbach, M. Rohrbach, S. Tang, S. Joon Oh, and B. Schiele, "Generating descriptions with grounded and co-referenced people," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4979–4989.

[25] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," The Journal of Machine Learning Research, vol. 21, no. 1, pp. 5485–5551, 2020.

[26] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 5659–5667.

[27] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10 578–10 587.

[28] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in International conference on machine learning. PMLR, 2021, pp. 10 347–10 357.

[29] A. A. Osman, M. A. W. Shalaby, M. M. Soliman, and K. M. Elsayed, "A survey on attention-based models for image captioning," International Journal of Advanced Computer Science and Applications, vol. 14, no. 2, 2023.

[30] R. Bernardi, R. Cakici, D. Elliott, A. Erdem, E. Erdem, N. Ikizler-Cinbis, F. Keller, A. Muscat, and B. Plank, "Automatic description generation from images: A survey of models, datasets, and evaluation measures," Journal of Artificial Intelligence Research, vol. 55, pp. 409–442, 2016.

[31] S. Bai and S. An, "A survey on automatic image caption generation," Neurocomputing, vol. 311, pp. 291–304, 2018.

[32] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," ACM Computing Surveys (CsUR), vol. 51, no. 6, pp. 1–36, 2019.

[33] X. Liu, Q. Xu, and N. Wang, "A survey on deep neural network-based image captioning," The Visual Computer, vol. 35, no. 3, pp. 445–470, 2019.

[34] H. Sharma, M. Agrahari, S. K. Singh, M. Firoj, and R. K. Mishra, "Image captioning: a comprehensive survey," in 2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC). IEEE, 2020, pp. 325–328.

[35] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt et al., "From captions to visual concepts and back," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1473–1482.

[36] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.

[37] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Advances in neural information processing systems, vol. 25, 2012.

[38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[39] J. Gu, G. Wang, J. Cai, and T. Chen, "An empirical study of language cnn for image captioning," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 1222–1231.

[40] X. Chen and C. Lawrence Zitnick, "Mind's eye: A recurrent visual representation for image caption generation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 2422–2431.

[41] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4651–4659.

[42] Q. Wu, C. Shen, L. Liu, A. Dick, and A. Van Den Hengel, "What value do explicit high level concepts have in vision to language problems?" in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 203–212.

[43] F. Chen, R. Ji, J. Su, Y. Wu, and Y. Wu, "Structcap: Structured semantic embedding for image captioning," in Proceedings of the 25th ACM international conference on Multimedia, 2017, pp. 46–54.

[44] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 7008–7024.

[45] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 375–383.

[46] B. Dai, D. Ye, and D. Lin, "Rethinking the form of latent states in image captioning," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 282–298.

[47] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv:1409.0473, 2014.

[48] X. Chen, L. Ma, W. Jiang, J. Yao, and W. Liu, "Regularizing rnns for caption generation by reconstructing the past with the present," in Proceedings of the IEEE Conference on computer vision and pattern recognition, 2018, pp. 7995–8003.

[49] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 4894–4902.

[50] Y. Wang, Z. Lin, X. Shen, S. Cohen, and G. W. Cottrell, "Skeleton key: Image captioning by skeleton-attribute decomposition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 7272–7281.

[51] H. Ge, Z. Yan, K. Zhang, M. Zhao, and L. Sun, "Exploring overall contextual information for image captioning in human-like cognitive style," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1754–1763.

[52] J. Gu, J. Cai, G. Wang, and T. Chen, "Stack-captioning: Coarse-to-fine learning for image captioning," in Proceedings of the AAAI conference on artificial intelligence, vol. 32, no. 1, 2018.

[53] Z. Yang, Y. Yuan, Y. Wu, W. W. Cohen, and R. R. Salakhutdinov, "Review networks for caption generation," Advances in neural information processing systems, vol. 29, 2016.

[54] W. Jiang, L. Ma, Y.-G. Jiang, W. Liu, and T. Zhang, "Recurrent fusion network for image captioning," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 499–515.

[55] Y. Sugano and A. Bulling, "Seeing with humans: Gaze-assisted neural image captioning," arXiv preprint arXiv:1608.05203, 2016.

[56] H. R. Tavakoli, R. Shetty, A. Borji, and J. Laaksonen, "Paying attention to descriptions generated by image captioning models," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2487–2496.

[57] V. Ramanishka, A. Das, J. Zhang, and K. Saenko, "Top-down visual saliency guided by captions," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 7206–7215.

[58] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Paying more attention to saliency: Image captioning with saliency and context attention," ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), vol. 14, no. 2, pp. 1–21, 2018.

[59] S. Chen and Q. Zhao, "Boosted attention: Leveraging human attention for image captioning," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 68–84.

[60] F. Chen, R. Ji, X. Sun, Y. Wu, and J. Su, "Groupcap: Group-based image captioning with structured relevance and diversity constraints," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 1345–1353.

[61] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[62] L. Ke, W. Pei, R. Li, X. Shen, and Y.-W. Tai, "Reflective decoding network for image captioning," in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 8888–8897.

[63] Y. Qin, J. Du, Y. Zhang, and H. Lu, "Look back and predict forward in image captioning," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 8367–8375.

[64] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," International journal of computer vision, vol. 123, pp. 32–73, 2017.

[65] M. Pedersoli, T. Lucas, C. Schmid, and J. Verbeek, "Areas of attention for image captioning," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 1242–1250.

[66] Z. Shi, X. Zhou, X. Qiu, and X. Zhu, "Improving image captioning with better use of captions," arXiv preprint arXiv:2006.11807, 2020.

[67] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-encoding scene graphs for image captioning," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 10 685–10 694.

[68] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning: Transforming objects into words," Advances in neural information processing systems, vol. 32, 2019.

[69] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer, 2014, pp. 740–755.

[70] M. Alikhani, P. Sharma, S. Li, R. Soricut, and M. Stone, "Clue: Cross-modal coherence modeling for caption generation," arXiv preprint arXiv:2005.00908, 2020.

[71] F. Liu, Y. Liu, X. Ren, X. He, and X. Sun, "Aligning visual regions and textual concepts for semantic-grounded image representations," Advances in Neural Information Processing Systems, vol. 32, 2019.

[72] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark et al., "Learning transferable visual models from natural language supervision," in International conference on machine learning. PMLR, 2021, pp. 8748–8763.

[73] J. Aneja, A. Deshpande, and A. G. Schwing, "Convolutional image captioning," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 5561–5570.

[74] L. Guo, J. Liu, J. Tang, J. Li, W. Luo, and H. Lu, "Aligning linguistic words and visual semantic units for image captioning," in Proceedings of the 27th ACM international conference on multimedia, 2019, pp. 765–773.

[75] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, "Guiding the long-short term memory model for image caption generation," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 2407–2415.

[76] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 2625–2634.

[77] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," Advances in neural information processing systems, vol. 28, 2015.

[78] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," arXiv preprint arXiv:1609.02907, 2016.

[79] T. Yao, Y. Pan, Y. Li, and T. Mei, "Hierarchy parsing for image captioning," in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 2621–2629.

[80] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.

[81] X. Yang, H. Zhang, and J. Cai, "Learning to collocate neural modules for image captioning," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 4250–4260.

[82] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever et al., "Improving language understanding by generative pre-training," 2018.

[83] L. Huang, W. Wang, Y. Xia, and J. Chen, "Adaptively aligned image captioning via adaptive attention time," Advances in neural information processing systems, vol. 32, 2019.

[84] L. Wang, Z. Bai, Y. Zhang, and H. Lu, "Show, recall, and tell: Image captioning with recall mechanism," in Proceedings of the AAAI conference on artificial intelligence, vol. 34, no. 07, 2020, pp. 12 176–12 183.

[85] Z.-J. Zha, D. Liu, H. Zhang, Y. Zhang, and F. Wu, "Context-aware visual policy network for fine-grained image captioning," IEEE transactions on pattern analysis and machine intelligence, vol. 44, no. 2, pp. 710–722, 2019.

[86] S. He, W. Liao, H. R. Tavakoli, M. Yang, B. Rosenhahn, and N. Pugeault, "Image captioning through image transformer," in Proceedings of the Asian conference on computer vision, 2020.

[87] J. Ji, Y. Luo, X. Sun, F. Chen, G. Luo, Y. Wu, Y. Gao, and R. Ji, "Improving image captioning by leveraging intra-and inter-layer global representation in transformer network," in Proceedings of the AAAI conference on artificial intelligence, vol. 35, no. 2, 2021, pp. 1655–1663.

[88] Z.-c. Fei, "Fast image caption generation with position alignment," arXiv preprint arXiv:1912.06365, 2019.

[89] G. Li, L. Zhu, P. Liu, and Y. Yang, "Entangled transformer for image captioning," in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 8928–8937.

[90] X. Hu, Z. Gan, J. Wang, Z. Yang, Z. Liu, Y. Lu, and L. Wang, "Scaling up vision-language pre-training for image captioning," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 17 980–17 989.

[91] Z. Fei, "Iterative back modification for faster image captioning," in Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 3182–3190.

[92] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 4634–4643.

[93] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-linear attention networks for image captioning," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10 971–10 980.

[94] H. Jiang, I. Misra, M. Rohrbach, E. Learned-Miller, and X. Chen, "In defense of grid features for visual question answering," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10 267–10 276.

[95] R. Mokady, A. Hertz, and A. H. Bermano, "Clipcap: Clip prefix for image captioning," arXiv preprint arXiv:2111.09734, 2021.

[96] L. Guo, J. Liu, X. Zhu, P. Yao, S. Lu, and H. Lu, "Normalized and geometry-aware self-attention network for image captioning," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10 327–10 336.

[97] F. Liu, X. Ren, X. Wu, S. Ge, W. Fan, Y. Zou, and X. Sun, "Prophet attention: Predicting attention with future attention," Advances in Neural Information Processing Systems, vol. 33, pp. 1865–1876, 2020.

[98] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[99] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao, "Vinvl: Revisiting visual representations in vision-language models," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 5579–5588.

[100] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[101] L. Guo, J. Liu, X. Zhu, and H. Lu, "Fast sequence generation with multi-agent reinforcement learning," arXiv preprint arXiv:2101.09698, 2021.
[102] P. Koehn, Statistical machine translation. Cambridge University Press, 2009.
[103] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," arXiv preprint arXiv:1511.06732, 2015.
[104] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," Machine learning, vol. 8, pp. 229–256, 1992.
[105] J. Lu, J. Yang, D. Batra, and D. Parikh, "Neural baby talk," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7219–7228.
[106] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," Advances in neural information processing systems, vol. 32, 2019.
[107] X. Yang, H. Zhang, D. Jin, Y. Liu, C.-H. Wu, J. Tan, D. Xie, J. Wang, and X. Wang, "Fashion captioning: Towards generating accurate descriptions with semantic rewards," in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16. Springer, 2020, pp. 1–17.
[108] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 49–58.
[109] A. F. Biten, L. Gomez, M. Rusinol, and D. Karatzas, "Good news, everyone! context driven entity-aware captioning for news images," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 12 466–12 475.
[110] L. A. Hendricks, R. Hu, T. Darrell, and Z. Akata, "Grounding visual explanations," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 264–279.
[111] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in International conference on machine learning. PMLR, 2016, pp. 1060–1069.
[112] J. Pont-Tuset, J. Uijlings, S. Changpinyo, R. Soricut, and V. Ferrari, "Connecting vision and language with localized narratives," in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16. Springer, 2020, pp. 647–664.
[113] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in International Conference on Machine Learning. PMLR, 2021, pp. 8821–8831.
[114] L. Guo, J. Liu, X. Zhu, X. He, J. Jiang, and H. Lu, "Non-autoregressive image captioning with counterfactuals-critical multi-agent learning," arXiv preprint arXiv:2005.04690, 2020.
[115] M. Cornia, L. Baraldi, G. Fiameni, and R. Cucchiara, "Universal captioner: Long-tail vision-and-language model training through content-style separation," arXiv preprint arXiv:2111.12727, vol. 1, no. 2, p. 4, 2021.
[116] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," arXiv preprint arXiv:1908.07490, 2019.
[117] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao, "Simvlm: Simple visual language model pretraining with weak supervision," arXiv preprint arXiv:2108.10904, 2021.
[118] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3558–3568.
[119] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-rnn)," arXiv preprint arXiv:1412.6632, 2014.
[120] D. Gurari, Y. Zhao, M. Zhang, and N. Bhattacharya, "Captioning images taken by people who are blind," in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16. Springer, 2020, pp. 417–434.
[121] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, and J. Gao, "Unified vision-language pre-training for image captioning and vqa," in Proceedings of the AAAI conference on artificial intelligence, vol. 34, no. 07, 2020, pp. 13 041–13 049.
[122] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona, "Caltech-ucsd birds 200. california institute of technology," CNS-TR-2010-001, Tech. Rep., 2010.
[123] S. Shen, L. H. Li, H. Tan, M. Bansal, A. Rohrbach, K.-W. Chang, Z. Yao, and K. Keutzer, "How much can clip benefit vision-and-language tasks?" arXiv preprint arXiv:2107.06383, 2021.
[124] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in 2008 Sixth Indian conference on computer vision, graphics & image processing. IEEE, 2008, pp. 722–729.
[125] A. Ramisa, F. Yan, F. Moreno-Noguer, and K. Mikolajczyk, "Breakingnews: Article annotation by image and text processing," IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 5, pp. 1072–1085, 2017.
[126] T.-H. Chen, Y.-H. Liao, C.-Y. Chuang, W.-T. Hsu, J. Fu, and M. Sun, "Show, adapt and tell: Adversarial training of cross-domain image captioner," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 521–530.
[127] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell, "Generating visual explanations," in Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14. Springer, 2016, pp. 3–19.
[128] O. Sidorov, R. Hu, M. Rohrbach, and A. Singh, "Textcaps: a dataset for image captioning with reading comprehension," in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. Springer, 2020, pp. 742–758.
[129] J. Kasai, K. Sakaguchi, L. Dunagan, J. Morrison, R. L. Bras, Y. Choi, and N. A. Smith, "Transparent human evaluation for image captioning," arXiv preprint arXiv:2111.08940, 2021.
[130] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005, pp. 65–72.
[131] N. Sharif, L. White, M. Bennamoun, and S. A. A. Shah, "Nneval: Neural network based evaluation metric for image captioning," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 37–53.
[132] Y. Cui, G. Yang, A. Veit, X. Huang, and S. Belongie, "Learning to evaluate image captioning," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 5804–5812.
[133] O. Caglayan, P. Madhyastha, and L. Specia, "Curious case of language generation evaluation metrics: A cautionary tale," arXiv preprint arXiv:2010.13588, 2020.
[134] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "Yfcc100m: The new data in multimedia research," Communications of the ACM, vol. 59, no. 2, pp. 64–73, 2016.
[135] K. Srinivasan, K. Raman, J. Chen, M. Bendersky, and M. Najork, "Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning," in Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 2443–2449.
[136] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in International conference on machine learning. PMLR, 2021, pp. 4904–4916.
[137] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, "Improved image captioning via policy gradient optimization of spider," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 873–881.
[138] R. Shetty, M. Rohrbach, L. Anne Hendricks, M. Fritz, and B. Schiele, "Speaking the same language: Matching machine to human captions by adversarial training," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 4135–4144.
[139] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14. Springer, 2016, pp. 382–398.
[140] J. Gao, S. Wang, S. Wang, S. Ma, and W. Gao, "Self-critical n-step training for image captioning," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 6300–6308.
[141] K. Desai and J. Johnson, "Virtex: Learning visual representations from textual annotations," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 11 162–11 173.
[142] A. Rohrbach, L. A. Hendricks, K. Burns, T. Darrell, and K. Saenko, "Object hallucination in image captioning," arXiv preprint arXiv:1809.02156, 2018.
[143] M. Jiang, J. Hu, Q. Huang, L. Zhang, J. Diesner, and J. Gao, "Reo-relevance, extraness, omission: A fine-grained evaluation for image captioning," arXiv preprint arXiv:1909.02217, 2019.
[144] Z. Wang, B. Feng, K. Narasimhan, and O. Russakovsky, "Towards unique and informative captioning of images," in Computer Vision–ECCV 2020:

16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16. Springer, 2020, pp. 629–644.

[145] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger, "From word embeddings to document distances," in International conference on machine learning. PMLR, 2015, pp. 957–966.

[146] Q. Wang, J. Wan, and A. B. Chan, "On diversity in image captioning: Metrics and methods," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 2, pp. 1035–1049, 2020.

[147] M. Kilickaya, A. Erdem, N. Ikizler-Cinbis, and E. Erdem, "Re-evaluating automatic metrics for image captioning," arXiv preprint arXiv:1612.07600, 2016.

[148] M. Cornia, L. Baraldi, and R. Cucchiara, "Show, control and tell: A framework for generating controllable and grounded captions," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 8307–8316.

[149] R. Bigazzi, F. Landi, M. Cornia, S. Cascianelli, L. Baraldi, and R. Cucchiara, "Explore and explain: self-supervised navigation and recounting," in 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021, pp. 1152–1159.

[150] H. Lee, S. Yoon, F. Dernoncourt, D. S. Kim, T. Bui, and K. Jung, "Vilbertscore: Evaluating image caption using vision-and-language bert," in Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems, 2020, pp. 34–39.

[151] Y. Yi, H. Deng, and J. Hu, "Improving image captioning evaluation by considering inter references variance," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 985–994.

[152] S. Wang, Z. Yao, R. Wang, Z. Wu, and X. Chen, "Faier: Fidelity and adequacy ensured image caption evaluation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14 050–14 059.

[153] Y. Luo, J. Ji, X. Sun, L. Cao, Y. Wu, F. Huang, C.-W. Lin, and R. Ji, "Dual-level collaborative transformer for image captioning," in Proceedings of the AAAI conference on artificial intelligence, vol. 35, no. 3, 2021, pp. 2286–2293.

[154] H. Lee, S. Yoon, F. Dernoncourt, T. Bui, and K. Jung, "Umic: An unreferenced metric for image captioning via contrastive learning," arXiv preprint arXiv:2106.14019, 2021.

[155] E. Van Miltenburg, D. Elliott, and P. Vossen, "Measuring the diversity of automatic image descriptions," in Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 1730–1741.

[156] Q. Wang and A. B. Chan, "Describing like humans: on diversity in image captioning," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4195–4203.

[157] B. Dai, S. Fidler, and D. Lin, "A neural compositional paradigm for image captioning," Advances in Neural Information Processing Systems, vol. 31, 2018.

[158] B. Dai, S. Fidler, R. Urtasun, and D. Lin, "Towards diverse and natural image descriptions via a conditional gan," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2970–2979.

[159] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," arXiv preprint arXiv:1904.09675, 2019.

[160] I. J. Unanue, J. Parnell, and M. Piccardi, "Berttune: Fine-tuning neural machine translation with bertscore," arXiv preprint arXiv:2106.02208, 2021.

[161] M. Jiang, Q. Huang, L. Zhang, X. Wang, P. Zhang, Z. Gan, J. Diesner, and J. Gao, "Tiger: Text-to-image grounding for image caption evaluation," arXiv preprint arXiv:1909.02050, 2019.

[162] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 201–216.

[163] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, "Clipscore: A reference-free evaluation metric for image captioning," arXiv preprint arXiv:2104.08718, 2021.

[164] L. A. Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, and T. Darrell, "Deep compositional captioning: Describing novel object categories without paired training data," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1–10.

[165] T. Yao, Y. Pan, Y. Li, and T. Mei, "Incorporating copying mechanism in image captioning for learning novel objects," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 6580–6588.

[166] S. Venugopalan, L. Anne Hendricks, M. Rohrbach, R. Mooney, T. Darrell, and K. Saenko, "Captioning images with diverse objects," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 5753–5761.

[167] H. Agrawal, K. Desai, Y. Wang, X. Chen, R. Jain, M. Johnson, D. Batra, D. Parikh, S. Lee, and P. Anderson, "Nocaps: Novel object captioning at scale," in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 8948–8957.

[168] Y. Li, T. Yao, Y. Pan, H. Chao, and T. Mei, "Pointing novel objects in image captioning," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 12 497–12 506.

[169] Y. Wu, L. Zhu, L. Jiang, and Y. Yang, "Decoupled novel object captioner," in Proceedings of the 26th ACM international conference on Multimedia, 2018, pp. 1029–1037.

[170] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Guided open vocabulary image captioning with constrained beam search," arXiv preprint arXiv:1612.00576, 2016.

[171] X. Hu, X. Yin, K. Lin, L. Zhang, J. Gao, L. Wang, and Z. Liu, "Vivo: Visual vocabulary pre-training for novel object captioning," in proceedings of the AAAI conference on artificial intelligence, vol. 35, no. 2, 2021, pp. 1575–1583.

[172] J. Gu, S. Joty, J. Cai, and G. Wang, "Unpaired image captioning by language pivoting," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 503–519.

[173] J. Gu, S. Joty, J. Cai, H. Zhao, X. Yang, and G. Wang, "Unpaired image captioning via scene graph alignments," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 10 323–10 332.

[174] D. Guo, Y. Wang, P. Song, and M. Wang, "Recurrent relational memory network for unsupervised image captioning," arXiv preprint arXiv:2006.13611, 2020.

[175] D.-J. Kim, J. Choi, T.-H. Oh, and I. S. Kweon, "Image captioning with very scarce supervised data: Adversarial semi-supervised learning approach," arXiv preprint arXiv:1909.02201, 2019.

[176] H. Ben, Y. Pan, Y. Li, T. Yao, R. Hong, M. Wang, and T. Mei, "Unpaired image captioning with semantic-constrained self-learning," IEEE Transactions on Multimedia, vol. 24, pp. 904–916, 2021.

[177] R. Del Chiaro, B. Twardowski, A. Bagdanov, and J. Van de Weijer, "Ratt: Recurrent attention to transient tasks for continual image captioning," Advances in Neural Information Processing Systems, vol. 33, pp. 16 736–16 748, 2020.

[178] J. Johnson, A. Karpathy, and L. Fei-Fei, "Densecap: Fully convolutional localization networks for dense captioning," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4565–4574.

[179] L. Yang, K. Tang, J. Yang, and L.-J. Li, "Dense captioning with joint inference and visual context," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2193–2202.

[180] X. Li, S. Jiang, and J. Han, "Learning object context for dense captioning," in Proceedings of the AAAI conference on artificial intelligence, vol. 33, no. 01, 2019, pp. 8650–8657.

[181] G. Yin, L. Sheng, B. Liu, N. Yu, X. Wang, and J. Shao, "Context and attribute grounded dense captioning," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 6241–6250.

[182] D.-J. Kim, J. Choi, T.-H. Oh, and I. S. Kweon, "Dense relational captioning: Triple-stream networks for relationship-based captioning," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 6271–6280.

[183] Y. Mao, C. Zhou, X. Wang, and R. Li, "Show and tell more: Topic-oriented multi-sentence image captioning." in IJCAI, 2018, pp. 4258–4264.

[184] M. Chatterjee and A. G. Schwing, "Diverse and coherent paragraph generation from images," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 729–744.

[185] J. Krause, J. Johnson, R. Krishna, and L. Fei-Fei, "A hierarchical approach for generating descriptive image paragraphs," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 317–325.

[186] X. Liang, Z. Hu, H. Zhang, C. Gan, and E. P. Xing, "Recurrent topic-transition gan for visual paragraph generation," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 3362–3371.

[187] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 684–699.

[188] Y. Luo, Z. Huang, Z. Zhang, Z. Wang, J. Li, and Y. Yang, "Curiosity-driven reinforcement learning for diverse visual paragraph generation," in Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 2341–2350.

[189] Z. Yang, Y. Lu, J. Wang, X. Yin, D. Florencio, L. Wang, C. Zhang, L. Zhang, and J. Luo, "Tap: Text-aware pre-training for text-vqa and text-caption," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 8751–8761.

[190] J. Wang, J. Tang, and J. Luo, "Multimodal attention with image text spatial relationship for ocr-based image captioning," in Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 4337–4345.

[191] J. Wang, J. Tang, M. Yang, X. Bai, and J. Luo, "Improving ocr-based image captioning by incorporating geometrical relationship," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 1306–1315.

[192] Q. Zhu, C. Gao, P. Wang, and Q. Wu, "Simple is not easy: A simple strong baseline for textvqa and textcaps," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 4, 2021, pp. 3608–3615.

[193] G. Xu, S. Niu, M. Tan, Y. Luo, Q. Du, and Q. Wu, "Towards accurate text-based image captioning with content diversity exploration," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 12 637–12 646.

[194] H. Jhamtani and T. Berg-Kirkpatrick, "Learning to describe differences between pairs of similar images," arXiv preprint arXiv:1808.10584, 2018.

[195] D. H. Park, T. Darrell, and A. Rohrbach, "Robust change captioning," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 4624–4633.

[196] X. Shi, X. Yang, J. Gu, S. Joty, and J. Cai, "Finding it at another side: A viewpoint-adapted matching encoder for change captioning," in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16. Springer, 2020, pp. 574–590.

[197] Q. Huang, Y. Liang, J. Wei, Y. Cai, H. Liang, H.-f. Leung, and Q. Li, "Image difference captioning with instance-level fine-grained feature representation," IEEE transactions on multimedia, vol. 24, pp. 2004–2017, 2021.

[198] H. Kim, J. Kim, H. Lee, H. Park, and G. Kim, "Agnostic change captioning with cycle consistency," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 2095–2104.

[199] M. Hosseinzadeh and Y. Wang, "Image change captioning by learning from an auxiliary task," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2725–2734.

[200] A. Vijayakumar, M. Cogswell, R. Selvaraju, Q. Sun, S. Lee, D. Crandall, and D. Batra, "Diverse beam search for improved description of complex scenes," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, 2018.

[201] B. Dai and D. Lin, "Contrastive learning for image captioning," Advances in Neural Information Processing Systems, vol. 30, 2017.

[202] L. Liu, J. Tang, X. Wan, and Z. Guo, "Generating diverse and descriptive image captions using visual paraphrases," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 4240–4249.

[203] L. Wang, A. Schwing, and S. Lazebnik, "Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space," Advances in Neural Information Processing Systems, vol. 30, 2017.

[204] J. Aneja, H. Agrawal, D. Batra, and A. Schwing, "Sequential latent spaces for modeling the intention during diverse image captioning," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 4261–4270.

[205] F. Chen, R. Ji, J. Ji, X. Sun, B. Zhang, X. Ge, Y. Wu, F. Huang, and Y. Wang, "Variational structured semantic inference for diverse image captioning," Advances in Neural Information Processing Systems, vol. 32, 2019.

[206] S. Mahajan and S. Roth, "Diverse image captioning with context-object split latent spaces," Advances in Neural Information Processing Systems, vol. 33, pp. 3613–3624, 2020.

[207] A. Deshpande, J. Aneja, L. Wang, A. G. Schwing, and D. Forsyth, "Fast, diverse and accurate image captioning guided by part-of-speech," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 10 695–10 704.

[208] D. Elliott, S. Frank, and E. Hasler, "Multilingual image description with neural sequence models," arXiv preprint arXiv:1510.04709, 2015.

[209] X. Li, C. Xu, X. Wang, W. Lan, Z. Jia, G. Yang, and J. Xu, "Coco-cn for cross-lingual image tagging, captioning, and retrieval," IEEE Transactions on Multimedia, vol. 21, no. 9, pp. 2347–2360, 2019.

[210] T. Miyazaki and N. Shimizu, "Cross-lingual image caption generation," in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, pp. 1780–1790.

[211] D. Elliott, S. Frank, K. Sima'an, and L. Specia, "Multi30k: Multilingual english-german image descriptions," arXiv preprint arXiv:1605.00459, 2016.

[212] W. Lan, X. Li, and J. Dong, "Fluency-guided cross-lingual image captioning," in Proceedings of the 25th ACM international conference on Multimedia, 2017, pp. 1549–1557.

[213] Y. Song, S. Chen, Y. Zhao, and Q. Jin, "Unpaired cross-lingual image caption generation with self-supervised rewards," in Proceedings of the 27th ACM international conference on multimedia, 2019, pp. 784–792.

[214] B. Jing, P. Xie, and E. Xing, "On the automatic generation of medical imaging reports," arXiv preprint arXiv:1711.08195, 2017.

[215] F. Liu, X. Wu, S. Ge, W. Fan, and Y. Zou, "Exploring and distilling posterior and prior knowledge for radiology report generation," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 13 753–13 762.

[216] X. Yang, M. Ye, Q. You, and F. Ma, "Writing by memorizing: Hierarchical retrieval-based medical report generation," arXiv preprint arXiv:2106.06471, 2021.

[217] Z. Bai, Y. Nakashima, and N. Garcia, "Explain me the painting: Multi-topic knowledgeable art description generation," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 5422–5432.

[218] Y. Feng and M. Lapata, "Automatic caption generation for news images," IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 4, pp. 797–812, 2012.

[219] A. Tran, A. Mathews, and L. Xie, "Transform and tell: Entity-aware news image captioning," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 13 035–13 045.

[220] F. Liu, Y. Wang, T. Wang, and V. Ordonez, "Visual news: Benchmark and challenges in news image captioning," arXiv preprint arXiv:2010.03743, 2020.

[221] X. Yang, S. Karaman, J. Tetreault, and A. Jaimes, "Journalistic guidelines aware news image captioning," arXiv preprint arXiv:2109.02865, 2021.

[222] S. Wu, J. Wieland, O. Farivar, and J. Schiller, "Automatic alt-text: Computer-generated image descriptions for blind users on a social network service," in proceedings of the 2017 ACM conference on computer supported cooperative work and social computing, 2017, pp. 1180–1192.

[223] C. C. Park, B. Kim, and G. Kim, "Towards personalized image captioning via multimodal memory networks," IEEE transactions on pattern analysis and machine intelligence, vol. 41, no. 4, pp. 999–1012, 2018.

[224] W. Zhang, Y. Ying, P. Lu, and H. Zha, "Learning long-and short-term user literal-preference with multimodal hierarchical transformer network for personalized image caption," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 05, 2020, pp. 9571–9578.

[225] C. Gan, Z. Gan, X. He, J. Gao, and L. Deng, "Stylenet: Generating attractive visual captions with styles," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 3137–3146.

[226] A. Mathews, L. Xie, and X. He, "Semstyle: Learning to generate stylised image captions using unaligned text," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8591–8600.

[227] L. Guo, J. Liu, P. Yao, J. Li, and H. Lu, "Mscap: Multi-style image captioning with unpaired stylized text," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4204–4213.

[228] W. Zhao, X. Wu, and X. Zhang, "Memcap: Memorizing style knowledge for image captioning," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 07, 2020, pp. 12 984–12 992.

[229] K. Shuster, S. Humeau, H. Hu, A. Bordes, and J. Weston, "Engaging image captioning via personality," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 12 516–12 526.

[230] Y. Zheng, Y. Li, and S. Wang, "Intention oriented image captions with guiding objects," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 8395–8404.

[231] Z. Meng, L. Yu, N. Zhang, T. L. Berg, B. Damavandi, V. Singh, and A. Bearman, "Connecting what to say with where to look by modeling human attention traces," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 12 679–12 688.
[232] L. Chen, Z. Jiang, J. Xiao, and W. Liu, "Human-like controllable image captioning with verb-specific semantic roles," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 16 846–16 856.
[233] S. Chen, Q. Jin, P. Wang, and Q. Wu, "Say as you wish: Fine-grained control of image caption generation with abstract scene graphs," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 9962–9971.
[234] Y. Zhong, L. Wang, J. Chen, D. Yu, and Y. Li, "Comprehensive image captioning via scene graph decomposition," in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16. Springer, 2020, pp. 211–229.
[235] C. Deng, N. Ding, M. Tan, and Q. Wu, "Length-controllable image captioning," in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16. Springer, 2020, pp. 712–729.
[236] F. Sammani and L. Melas-Kyriazi, "Show, edit and tell: a framework for editing image captions," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 4808–4816.
[237] M. Hodosh and J. Hockenmaier, "Focused evaluation for image description with binary forced-choice tasks," in Proceedings of the 5th Workshop on Vision and Language, 2016, pp. 19–28.
[238] H. Xie, T. Sherborne, A. Kuhnle, and A. Copestake, "Going beneath the surface: Evaluating image captioning for grammaticality, truthfulness and diversity," arXiv preprint arXiv:1912.08960, 2019.
[239] I. Lasri, A. Riadsolh, and M. Elbelkacemi, "Facial emotion recognition of deaf and hard-of-hearing students for engagement detection using deep learning," Education and Information Technologies, vol. 28, no. 4, pp. 4069–4092, 2023.
[240] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Universal image-text representation learning," in European conference on computer vision. Springer, 2020, pp. 104–120.
[241] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in International conference on machine learning. PMLR, 2017, pp. 3319–3328.

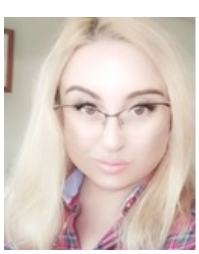
AZHAR JAMIL is currently pursuing a MS degree from Barani Institute of Information Technology, PMAS-Arid Agriculture University, Rawalpindi, Pakistan. His research interests include Artificial Intelligence, Neuroscience, Behavioral Data Modeling & Prescriptive Analytics, Data Visualization, Explority Data Analytics, Big Data Solution Engineering, Data Mining, Opinion Mining and Process Mining.

SAIF UR REHMAN is currently working as an Assistant Professor at University Institute of Information Technology, PMAS-Arid Agriculture University Rawalpindi, Pakistan. His research interests have been Artificial Intelligence, Machine Learning Data Mining, Graph Mining and Social Network Analysis.

KHALID MAHMOOD received his PhD degree in Computer Science in 2020 from Gomal University, D.I.Khan, Pakistan. Currently he is working as a faculty member in Institute of Computing and Information Technology (ICIT), Gomal University, D.I.Khan. His research interests are focused around Machine Learning, Deep Learning, Sentiment Analysis and Opinion Mining, Algorithms, and Information Security.

MONICA GRACIA VILLAR is working at Universidad Europea del Atlántico. Isabel Torres 21, 39011 Santander, Spain. She is also affiliated with Universidade Internacional do Cuanza. Cuito, Bié, Angola and Universidad de La Romana. La Romana, República Dominicana

THOMAS PROLA is working at Universidad Europea del Atlántico. Isabel Torres 21, 39011 Santander, Spain. He is also affiliated with Universidad Internacional Iberoamericana Campeche 24560, México and Universidad Internacional Iberoamericana Arecibo, Puerto Rico 00613, USA.



ISABEL DE LA TORRE-DÍEZ is currently a Professor with the Department of Signal Theory and Communications and Telematic Engineering, University of Valladolid, Spain, where she is also the Leader of the GTe Research Group (http://sigte.tel.uva.es). Her research interests include design, development, and evaluation of telemedicine applications, services and systems, e-health, m-health, electronic health records (EHRs), EHRs standards, biosensors, cloud and fog computing, data mining, quality of service (QoS), and quality of experience (QoE) applied to the health field.

MD ABDUS SAMAD (Member, IEEE) received the Ph.D. degree in information and communication engineering from Chosun University, South Korea. He worked as an Assistant Professor at the Department of Electronics and Telecommunication Engineering, International Islamic University Chittagong, Chattogram, Bangladesh, from 2013 to 2017. He has been working as a research professor at the Department of Information and Communication Engineering at Yeungnam University, South Korea. His research interests include signal processing, antenna design, electromagnetic wave propagation, applications of artificial neural networks, and millimeter-wave propagation by interference and atmospheric causes for 5G and beyond wireless networks. He won the Prestigious Korean Government Scholarship (GKS) for his doctoral study.

IMRAN ASHRAF received his Ph.D. in Information and Communication Engineering from Yeungnam University, South Korea in 2018, and the M.S. degree in computer science from the Blekinge Institute of Technology, Karlskrona, Sweden, in 2010 with distinction. He has worked as a postdoctoral fellow at Yeungnam University, as well. He is currently working as an Assistant Professor at the Information and Communication Engineering Department, Yeungnam University, Gyeongsan, South Korea. His research areas include positioning using next-generation networks, communication in 5G and beyond, location-based services in wireless communication, smart sensors (LIDAR) for smart cars, and data analytics.