

Attention-Based Transformer Models for Image Captioning Across Languages: An In-depth Survey and Evaluation

Israa A. Albadarneh^a, Bassam H. Hammo^{a,b}, Omar S. Al-Kadi*^a

^a*The University of Jordan, Amman, 11941, Jordan*

^b*Princess Sumaya University for Technology, Amman, 11941, Jordan*

Abstract

Image captioning involves generating textual descriptions from input images, bridging the gap between computer vision and natural language processing. Recent advancements in transformer-based models have significantly improved caption generation by leveraging attention mechanisms for better scene understanding. While various surveys have explored deep learning-based approaches for image captioning, few have comprehensively analyzed attention-based transformer models across multiple languages. This survey reviews attention-based image captioning models, categorizing them into transformer-based, deep learning-based, and hybrid approaches. It explores benchmark datasets, discusses evaluation metrics such as BLEU, METEOR, CIDEr, and ROUGE, and highlights challenges in multilingual captioning. Additionally, this paper identifies key limitations in current models, including semantic inconsistencies, data scarcity in non-English languages, and limitations in reasoning ability. Finally, we outline future research directions, such as multimodal learning, real-time applications in AI-powered assistants, healthcare, and forensic analysis. This survey serves as a comprehensive reference for researchers aiming to advance the field of attention-based image captioning.

Keywords: Image Captioning, Transformer, Attention Mechanism, Convolutional Neural Network, Computer Vision, Natural Language Processing

1. Introduction

Image captioning involves generating an image description, which includes identifying important objects and their relationships and creating syntactically and semantically correct sentences. This task requires collaboration between the computer vision (CV) and natural language processing (NLP) research communities [1] [2]. The large volume of unannotated images on the Internet has driven the automated image captioning process [3]. Furthermore, advances in deep learning models have significantly improved computer vision and natural language processing capabilities [4] [5]. Fig.1 shows the general architecture of the image captioning model.

The process of image captioning starts with an input image. The next step is image processing, which involves resizing, normalizing, and augmenting the image. This is followed by feature extraction using CNN architectures like ResNet or Inception, which encode the features into a fixed-size vector. In the language processing stage, models like RNN, LSTM, GRU, or Transformer convert words into vectors and predict the next word in the sequence. The attention mechanism selectively focuses on different parts of the image to enhance the captioning process. Finally, in the output stage, a descriptive caption is generated. For example, "A little girl in a pink dress going into a wooden cabin." This architecture effectively combines computer vision and natural language processing to automatically

generate descriptive text for images.

The field of computer vision has made significant progress in tasks such as object recognition, image segmentation, image classification, and scene recognition. However, generating a natural language description of an image using a computer is generally a complex task [6].

Image captioning combines research from the computer vision and natural language processing communities [1]. A captioning model aims to present a scene and text, a task that comes naturally to the human brain. It is common for humans to interpret information quickly from an image at one glance. However, challenges such as parallax errors make image captioning a complex problem that is not fully resolved. This error can make it difficult for the human eye and computer vision systems to detect objects at certain angles where their appearance changes, making them hard to recognize. In addition, objects of the same class might have various shapes and appearances from different angles, further complicating the task. Overlapping objects and scene clutter also pose challenges for accurate object detection [7].

Image captioning approaches have three main categories: template-based, retrieval-based, and deep learning-based. Template-based techniques use predefined templates with a set number of blank slots to generate captions. These approaches first recognize different objects, characteristics, and actions and then fill in the blank spaces in the templates. Although this approach can provide grammatically correct captions and relevant descriptions, the coverage, inventiveness, and complexity of the generated sentences coverage, inventiveness, and complexity

*Corresponding author

Email address: o.alkadi@ju.edu.jo (Omar S. Al-Kadi*)

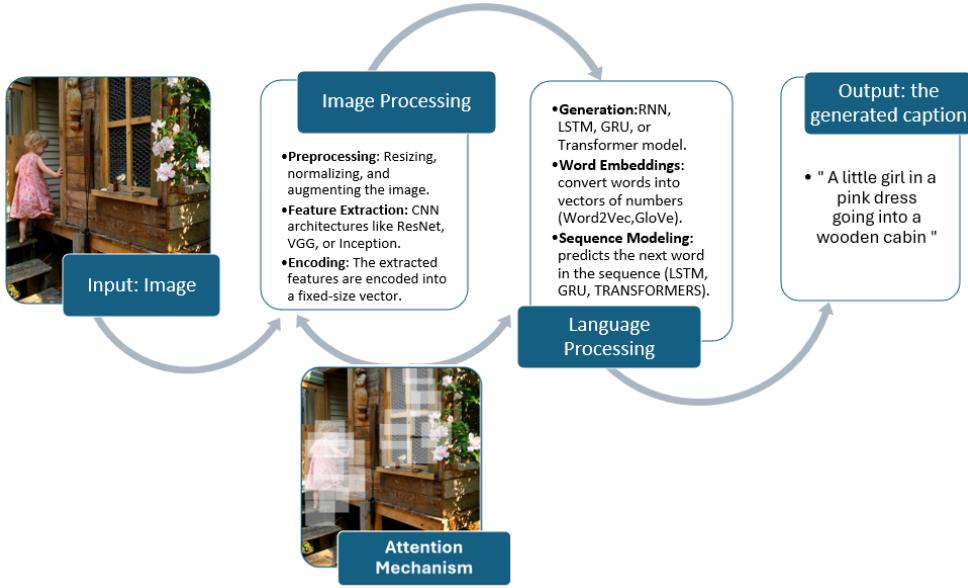


Figure 1: A general architecture for image captioning models.

are limited. In retrieval-based systems, captions are retrieved from a set of existing captions. This approach has the advantage of generating generic and syntactically correct captions. However, it may not produce semantically accurate captions specific to the image. Both template-based and retrieval-based approaches are not flexible enough, as they rely on existing captions in the training set or hard-coded language structures [2].

Given the challenges of using template-based and retrieval-based approaches, a third approach based on deep learning has been introduced. This approach follows recent developments in deep neural networks, widely used in computer vision and natural language processing. Deep neural networks can provide effective solutions for visual and language modeling [6]. As a result, they have been used to improve existing systems and create many innovative approaches [1]. Before significant advances in deep learning methods, image captioning was done mainly using traditional machine learning-based techniques, which included feature extraction methods like Scale-Invariant Feature Transform (SIFT), Local Binary Patterns (LBP), and Histogram of Oriented Gradients (HOG) [8, 9]. A classifier was then used to classify the items after feature extraction. However, traditional approaches are less preferred than deep learning-based methods, which automatically learn features due to the complexity of feature extraction from massive amounts of data [7]. Numerous recent publications have focused on applying deep machine learning to caption images [2].

Deep learning algorithms are effective at handling the complexities of image captioning. This survey offers a comprehensive analysis of transformer models and attention mechanisms in image captioning, providing a current overview of the relevant literature. The discussion is structured around the following research questions.

a) **RQ1:** What are the recent advancements and current status of image captioning models, particularly attention-based

and transformer-based models?

- b) **RQ2:** How have transformer-based models advanced image captioning across different languages, and what challenges and limitations do these models face?
- c) **RQ3:** How is image captioning applied across various domains, and what key evaluation metrics are used to assess the quality of generated captions?
- d) **RQ4:** What are the key conclusions from recent surveys on image captioning, and how do they guide future research directions?

1.1. Application of image captioning

Research in image captioning has gained significant importance across various fields. With its broad applicability and growing technological advancements, image captioning drives innovation across numerous domains in multiple real-world applications, offering significant benefits to industry and society.

In the field of medical imaging, for instance, surgeons can apply image captioning to monitor therapy progress preoperatively, intraoperatively, and postoperatively [10]. In education, researchers are exploring e-learning systems that integrate image captioning to enhance web-based learning experiences [11][12]. For visually impaired individuals, transformer-based photo captioning frameworks are being developed to translate visual content into written and spoken descriptions, improving accessibility and promoting self-reliance [13]. Similarly, specialized image captioning models in smart local tourism are expected to power AI-driven platforms, providing enriched experiences for travelers and local businesses alike [14].

Beyond these applications, image captioning plays a crucial role in virtual assistants [15], image retrieval [16], and information retrieval [17], as well as enhancing user engagement on social media platforms [18]. Additionally, researchers are exploring its potential in emerging technologies such as automated

self-driving cars [19], CCTV footage analysis [20], improving image search accuracy [21], and enhanced facial recognition systems [22].

1.2. Contribution and scope

Several survey articles have been published on the subject of image captioning. Although these surveys provided a good overview of the literature on image captioning, they did not cover publications discussing image captioning techniques for various languages. In addition, new deep-learning studies have been published since the survey papers were written. The key contributions of this survey are: (a) providing a comprehensive review of the current state of image captioning for various languages, specifically attention-based models, (b) discussing the detailed design of transformer models with different attention mechanisms, and (c) addressing ongoing challenges and highlighting potential future directions for the field. To highlight the unique contributions of this survey, Table 1 compares our study with previous ones on attention-based image captioning. Unlike prior reviews, this survey extensively covers multilingual models, a broader dataset range, and an in-depth evaluation of current challenges and future directions.

1.3. Search criteria

The survey included papers published between 2018 and 2024. A keyword search was conducted on Google Scholar using the following terms: image captioning, image description, image text generation, transformer-based image captioning, and attention-based image captioning. Google Scholar was chosen to avoid bias towards any specific publisher [30]. The survey covers articles on image captioning in multiple languages, including, but not limited to, English, Arabic, Vietnamese, Myanmar, and Indonesian.

1.4. Survey structure

This paper is organized as follows: Section 2 provides an overview of related surveys in image captioning, summarizes them, and discusses their foundations. Section 3 discusses the methods employed in image captioning models. Section 4 describes relevant research in image captioning, classifying them into five categories: handcrafted approaches, deep learning for image captioning, transformer-based approaches, attention-based approaches, and graph-based representation. Section 5 discusses the datasets. Section 6 introduces the evaluation metrics used to assess the quality and performance of generated captions. Limitations and challenges are summarized in Section 7, and finally, conclusions and future directions are presented in Section 8. Fig. 15 illustrates the structure of this survey.

2. Overview of Recent Surveys

Several recent papers have used deep learning techniques to create captions for images. This section presents a summary and analysis of relevant surveys in image captioning.

2.1. Attentive deep learning models

A literature survey by [31] demonstrated that bottom-up attention models, which combine multi-head attention, yield the most significant results. [23] proposed an attention-based deep learning model for image captioning as part of a comparative study. This research focused on attention mechanisms and identified key image areas based on the image's context, noting that attention can be beneficial in generating image captions. [24] reviewed advanced captioning techniques and classified them into attentive, semantically enhanced, transformation-based, post-editing, and vision-language pre-training (VLP).

The review by [32] examines the advancements in image captioning, tracing its evolution from traditional ML techniques to modern deep learning-based approaches. The study introduces a structured taxonomy for classifying image captioning methodologies and highlights key developments, including template-based, retrieval-based, and encoder-decoder models. Despite significant progress, the authors emphasize that further research is needed to develop more reliable and adaptable models. Similarly, [33] explores the evolution and persistent challenges of image captioning across various application domains, such as multimodal search engines, security, remote sensing, medical imaging, and assistive technologies. The study underscores the ongoing difficulties in achieving real-time captioning in critical fields like healthcare and security and the limited availability of large, domain-specific datasets. Additionally, issues related to training and evaluation continue to pose obstacles. While significant advancements have been made, the authors stress the need for continued research to enhance the robustness and practicality of image captioning models. In another survey by [34], the authors focus specifically on attention-based image captioning, reviewing the major breakthroughs in this area. The paper presents a new taxonomy for classifying attention-based techniques and discusses the challenges that hinder further development.

2.2. Segmentation and semantic analysis models

The application of deep learning approaches to segmentation analysis of 2D and 3D images was discussed in a study by [35]. The study highlighted that while applying deep learning methods to segmentation analysis might seem straightforward for humans, it remains a challenge for computers due in part to the limited understanding of the functioning and processing mechanisms of the human brain. The study categorized supervised learning-based techniques into encoder-decoder architecture-based, compositional architecture-based, attention-based, semantic concept-based, stylized captions, dense image captioning, and novel object-based image captioning, as outlined by [2].

2.3. Classical model

The study by [25] aimed to identify major technical advances in architectures and training methods and to analyze various relevant state-of-the-art methodologies. This work served as a valuable resource for understanding the existing literature

Table 1: Comparison of this survey with existing reviews on attention-based image captioning

Survey	Year	Focus	Attention-Based Models	Multilingual Coverage	Datasets Reviewed	Evaluation Metrics
[23]	2019	Comparative study of deep learning models	✓	✗	Limited (MS, COCO, Flickr8k)	BLEU, METEOR
[24]	2021	Transformer-based models	✓	✗	Large-scale datasets	CIDEr, SPICE
[25]	2022	Vision-language models	✓	✗	MS COCO, Flickr30k	Multiple metrics
[26]	2022	Review of datasets and metrics	✓	✗	Extensive dataset coverage	Detailed metric analysis
[27]	2021	Attentive deep learning models for IC	✓	✗	MS-COCO	B4, METEOR, CIDEr and SPICE
[28]	2024	Graph types used in 2D image understanding approaches	✓	✗	A summary of common datasets	A summary of performance metrics
[29]	2023	Medical image captioning	✓	✗	Common data set of medical IC	Multiple metrics
This Survey	2025	Comprehensive review of attention-based models, multilingual coverage	✓	✓ (English, Arabic, Vietnamese, etc.)	Extensive dataset coverage	Detailed metric analysis

and outlining potential future possibilities. In addition, a systematic review of the literature (SLR) summarized advances in image captioning [36]. The primary objective of this research was to summarize the findings from recent papers and to describe the most popular methods and challenging problems in image captioning.

The survey [37] explores the challenges and advancements in image captioning. The frameworks traditionally relied on a two-step pipeline, where visual features were extracted before being processed into natural language descriptions. However, with the emergence of sequential deep learning models, such as Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Gated Recurrent Units (GRUs), the efficiency and accuracy of caption generation have improved significantly. The paper provides a review of the modeling architectures used, and highlights key research challenges.

2.4. *Taxonomy of visual encoding and language modeling techniques*

The purpose of the [26] study is to present a thorough review of image captioning techniques. The researcher created a taxonomy of visual encoding and language modeling techniques, emphasizing their essential features and restrictions. A comprehensive survey was presented in [2]. This survey report provided an analysis of current deep learning-based image captioning methods, and a taxonomy of image captioning methods was provided. This study concluded that although deep learning-based image captioning techniques have made significant strides in recent years, a reliable technique still needs to be developed.

The survey by [28] examines the role of graph neural networks (GNNs) in 2D image understanding, a challenging problem in computer vision that aims to achieve human-level scene comprehension. Graphs are widely used in 2D image under-

standing pipelines as they effectively represent the relational structure between objects in an image. The study provides a detailed taxonomy of graph types employed in this domain, an extensive review of GNN models applied to 2D image understanding, and a forward-looking roadmap for future research. This survey covers key applications such as image captioning, visual question answering, and image retrieval, specifically focusing on approaches that exploit GNN-based architectures.

2.5. Interpretability of deep neural networks model

The question of the interpretability of deep neural networks, especially in image captioning methods based on classification or classification, was discussed in a survey by [38]. Due to the highly nonlinear functions and ambiguous working mechanisms, many works have aimed to explain the characteristics of ‘black box’ models. As deep learning models are often considered black boxes, the survey conducted by [39] aims to assess the impact of each module to enhance our understanding of the model. This research conducted quantitative and qualitative analyses to study the effects of five modules: the sequential module, the word embedding module, the initial seed module, the attention module, and the search module.

2.6. Transformer-based model

Several studies have explored the use of deep machine learning in image captioning, with a focus on transformer-based attention algorithms. However, there is a lack of thorough investigation into using transformer-based methodologies in image captioning. This gap in existing image captioning surveys has inspired this work to provide a comprehensive review of transformer-based approaches in image captioning, particularly focusing on attention-based methods. The attention mechanism for generating image captions is an area of increasing research interest due to its consistent relevance. Initial attempts to address this issue using transformer-based approaches have shown exceptional performance. Therefore, employing transformers to improve image captioning holds great promise [26].

2.7. Medical imaging reports

The study by [29] explores prospective advancements in the automatic generation of medical imaging reports using deep learning. Inspired by image captioning techniques, deep learning algorithms have significantly improved the efficiency and accuracy of diagnostic report generation. The paper provides a review of research efforts in this domain, focusing on deep learning architectures such as hierarchical RNN-based frameworks, attention-based models, and reinforcement learning-based approaches. Additionally, it examines the applications, underlying architectures, datasets, and evaluation methods used in medical imaging report generation. The study identifies key challenges in the field and proposes future research directions to enhance clinical applications and decision-making through more advanced report-generation methods.

Similarly, [40] presents an analysis of transformer networks in computer vision, with a special emphasis on their applications in natural and medical image analysis. Although transformers were initially designed for natural language processing, their recent adaptation to image-based tasks has demonstrated promising results, positioning them as a viable alternative to traditional convolutional neural networks. The review highlights core principles of the transformer’s attention mechanism, which enables effective long-range feature extraction. It explores various transformer-based architectures applied to critical tasks such as image segmentation, classification, registration, and diagnosis. The paper also highlights the current limitations of transformer networks in image analysis and outlines potential research directions to enhance their effectiveness, particularly in the context of medical imaging.

Expanding on this, [41] investigates the application of Vision Transformers (ViTs) in the medical domain. Initially inspired by the success of Transformer networks in language processing, ViTs have emerged as a powerful alternative to CNNs for computer vision tasks. These models and their variants excel at capturing long-range dependencies and spatial correlations, offering substantial benefits for medical image analysis tasks, including classification, segmentation, registration, detection, and radiological report generation. The paper specifically discusses the role of transformers in medical image captioning and disease diagnosis, providing insights into commonly used medical imaging modalities in clinical practice. Furthermore, it reviews the self-attention mechanism in vision transformers as applied to disease diagnosis and automated report generation. The study concludes by identifying existing challenges in the field and suggesting potential future research directions to enhance the efficiency of AI-driven applications in healthcare.

2.8. Remote sensing image captioning

The study by [42] explores the emerging field of remote sensing image captioning, which focuses on automatically generating textual descriptions for images captured by satellites, aircraft, and drones. As an interdisciplinary task integrating computer vision and natural language processing, remote sensing image captioning has garnered significant research interest in recent years. The paper analyzes relevant articles, summarizing key technical approaches, datasets, evaluation metrics, and experimental findings from state-of-the-art methods. Additionally, it examines the field’s strengths, limitations, and ongoing challenges while proposing valuable directions for future research. Similarly, [43] investigates the challenge of generating precise and adaptable textual descriptions for remote sensing images. While significant progress has been made in related tasks such as object detection and scene classification, accurately and concisely describing remote sensing imagery remains a complex problem. To address this issue, the paper introduces a set of annotation guidelines tailored to the unique characteristics of remote-sensing images, aiming to improve captioning quality. Additionally, the authors present a large-scale aerial image dataset specifically designed for remote sensing image captioning. Extensive experiments on this dataset

demonstrate that the generated English descriptions effectively capture the content of remote-sensing images.

3. Image Captioning Methods

This section discusses various methods in image captioning models, including deep learning-based, transformer-based, and attention-based approaches.

3.1. Deep learning-based approaches

The creation of image captions or descriptions can be approached in various ways. Common architectures such as CNN, RNN, and LSTM are often used to generate image captions. A convolutional neural network (CNN), an artificial intelligence (AI) network, has been utilized in many fields, including pattern recognition and natural language processing. Artificial neural networks (ANNs) are mathematical models with layers typically consisting of an input layer, an output layer, and one or more hidden layers. If x represents the input and f is the activation function, mathematically, a neuron can be represented as

$$z = f\left(\sum_{i=0}^n w_i x_i + \beta\right) \quad (1)$$

where n is the number of input features, w is the connection weights between the input layer and the hidden layer, β is the bias weight.

In most deep learning models, CNN is an encoder network, while RNNs are used as language-model decoder networks. However, some image captioning models use RNN for the encoder and decoder networks. A recurrent neural network includes an LSTM (long-short-term memory) component for long-term and short-term memory. LSTM is used for sentence representation to create image captions and extract features of images and words [23]. However, RNNs, LSTMs, and GRUs are susceptible to problems such as vanishing gradients, training difficulties, and long sequences. RNNs may not retain all information at the beginning of a long sequence. The specific operations of the LSTM-based decoder used in [6] to generate captions are described in (2), (3), and (4).

$$\begin{bmatrix} i_t \\ f_t \\ o_t \\ g_t \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{bmatrix} T_{D+m+n,n} \begin{bmatrix} E_{y_{(t-1)}} \\ h_{t-1} \\ \hat{z}_t \end{bmatrix} \quad (2)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \quad (3)$$

$$h_t = o_t \odot \tanh(c_t). \quad (4)$$

Input, forget, memory, output gates, and hidden state are represented by variables i_t , f_t , c_t , o_t , and h_t , respectively. T is a mapping with the formula $f_{s,t} : \mathbb{R}^s \rightarrow \mathbb{R}^t$. As a result, $\mathbb{R}^{(D+m+n)}$ to \mathbb{R}^n is mapped by $T_{D+m+n,n}$. $\hat{z} \in \mathbb{R}^D$ stands for the context vector that captures the visual data of a particular area in the input

image. E stands for the embedding matrix of dimension $m \times k$. The dimension of the embedding vector is indicated by the letter m , and the letter n indicates the dimension of the hidden state LSTM. Furthermore, σ and \odot represent logistic sigmoid and element-wise multiplication, respectively. A typical LSTM unit is shown in Fig.2.

The Long Short-Term Memory Network (LSTM) is a type of recurrent neural network (RNN) known for its superior performance. However, training LSTM networks can be challenging due to the complex addressing and overwriting mechanisms, the inherently sequential nature of the required processing, and the significant amount of storage needed during the procedure [7]. While LSTMs are slower at processing than CNNs, they excel at modeling dynamic temporal behavior in language, which cannot be achieved using only a language model [2]. On the other hand, global CNN features are known for their ease of use and compact representation. However, this approach also leads to excessive information compression and requires granularity, making it difficult for a captioning model to provide detailed descriptions [25].

In machine learning, another technique is reinforcement learning, while unsupervised learning methods include generative adversarial networks (GANs). GAN-based image captioning systems are capable of producing a variety of image descriptions. However, text processing relies on discrete numbers, making the processes non-differentiable and challenging to apply back-propagation directly. The architecture of the method presented by [44] is shown in Fig.3. It uses a GAN-based model to generate artificial images from text, employing attention to focus on relevant word vectors to create various parts of the image. Subsequently, captions are produced for the image using an attention-based image captioning model. [45] introduced a Gated Recurrent Unit (GRU) based on the generative adversarial structure network (GASN), which consists of three parts: a consensus reasoning module, a sentence decoder with two layers of LSTM, and a grounding module to locate regions. This method provided accurate and detailed information on objects to predict words.

3.2. Transformer-based approaches

The transformer is a neural network architecture introduced in [46]. It excels at handling sequential text data and comprises a stack of encoder and decoder layers. Each encoder and decoder stack contains the corresponding embedding layers for their inputs and an output layer to generate the final output. The encoder includes a self-attention layer for calculating relationships between words in the sequence, a feedforward layer, and a second encoder-decoder attention layer. Residual skip connections and two LayerNorm layers surround the encoder and decoder layers. Data inputs for the encoder and decoder include the embedding and position encoding layers. The encoder stack consists of multiple encoders, each with a feedforward and multi-head attention layers. In contrast, the decoder stack includes multiple decoders, each with two feedforward layers and multi-head attention [46].

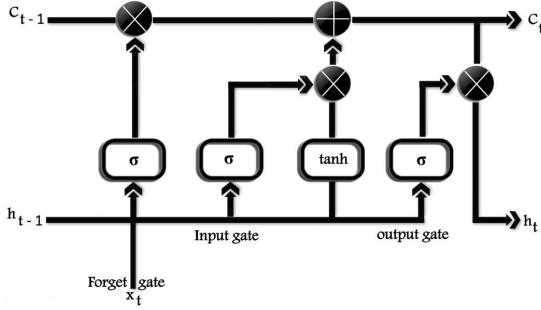


Figure 2: A typical LSTM unit consisting of forget, input, and output gates.

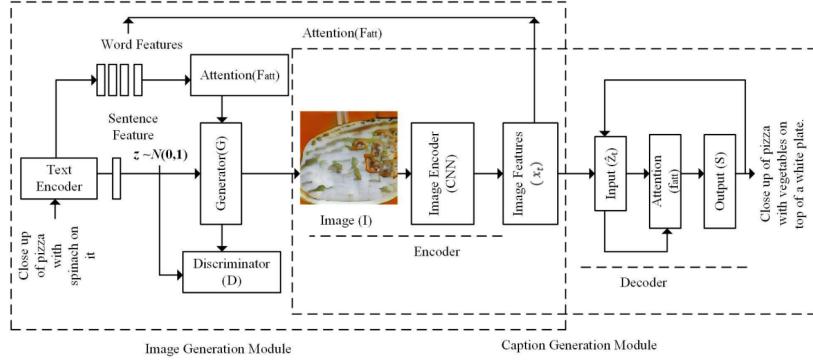


Figure 3: GAN-based model for image captioning [44].

Recent image captioning models leverage transformer architectures to connect informative regions in the image using attention, resulting in excellent performance. However, some previous transformer-based image captioning models have limitations because the transformer’s internal architecture was originally designed for machine translation. Text sequences are inherently sequential, whereas images are two- or three-dimensional, leading to significant differences in the relative spatial relationships between regions in images compared to phrases [47].

The transformer consists of two main parts: an encoder and a decoder. Multi-head attention functions as parallel heads of self-attention. Self-attention is the mechanism used by transformers to incorporate the context of other relevant words into the processing of the current word. Another component is the fully connected feedforward network, consisting of two linear transformations consistent across positions but varying parameters from layer to layer. The transformer adds a vector to each input embedding to help determine the position of each word; position embedding is a way of considering the order of words in an input sequence. The linear layer is a simple, fully connected neural network that transforms the vector produced by the decoder stack into a much larger vector known as a logit vector. SoftMax provides the probabilities. The cell with the highest probability is chosen, and the word associated with it is produced as the output [46].

The transformer model addresses the limitations of RNN and LSTM by enabling more parallelization and improving trans-

lation quality. Unlike RNN or LSTM, which process sentences one word at a time, transformer models can handle complete sentences through attention-based mechanisms [48]. Although RNN has challenges in scaling to larger levels, attention-guided image captioning can outperform later transformer-based techniques when used with strong visual encoders. Although these methods are often smaller than transformer-based approaches, they require longer training times. The transformer-based approach resolves the issue of long-distance dependency present in RNN. Furthermore, its structure makes it easier to scale the transformer model to deeper levels following the actual design requirements [26].

3.2.1. The transformer model

The transformer network uses an encoder-decoder architecture similar to RNN but with a key distinction. Unlike RNNs, transformers can simultaneously process the entire input sentence or sequence without any time step associated with the input. Transformers consist of N identical layers, each containing three sub-layers. The first layer utilizes a multi-head self-attention technique, including a mechanism to prevent the model from seeing future data, ensuring that the model only uses prior words to generate the current term. The second layer performs multi-head attention over the output of the first layer, serving as the foundation for correlating text and visual information with the attention mechanism. The third layer is a fully connected feedforward network. Following layer normaliza-

tion, the transformer applies a residual connection around the three sub-layers. Unlike LSTMs, the transformer can process all words in the caption simultaneously.

Transformers do not rely on recurrence or convolution and thus need to learn the relative or absolute positions of the words in a sequence. This is achieved by employing learned weights that represent the position of a token within a sentence. The fully attentive paradigm proposed by [46] has significantly transformed the way language production is viewed, leading the Transformer model to become the cornerstone of many NLP innovations and the *de facto* standard architecture for numerous language processing tasks.

The Transformer design has been utilized for image captioning, which can be considered a sequence-to-sequence task. In the conventional transformer decoder, words undergo a masked self-attention operation, followed by a cross-attention operation where words act as queries, and the output of the final encoder layer acts as keys and values, along with a final feedforward network. During training, a masking strategy is used to limit the influence of the preceding words [25]. Both the encoder and decoder of the Transformer utilize layered self-attention and point-wise interconnected layers, as shown in the left and right halves of Fig.4. Self-attention, or intra-attention, focuses on the relationships between different positions in a single sequence to represent the sequence. Self-attention has been successfully applied in reading comprehension, abstractive summarization, textual entailment, and sentence representations independent of the learning task [46].

The attention mechanism focuses on a subset of the details relevant to our objective instead of assessing the entire picture simultaneously. The core of the attention mechanism lies in selecting the portion of detail to concentrate on based on our goals and continually analyzing it. By calculating the similarity of word vectors, self-attention determines the degree of correlation between the current word and other words for the image captioning task. Typically, two-word vectors have smaller distance angles and greater products the closer their meanings are to each other. By normalizing the similarity, weights are generated. The attention score also referred to as the level of attention of the current word to other words, is obtained by multiplying the weights by the word vectors and summing them. The feed-forward network, a one-way propagation neural network, can be classified based on the sequence in which information is received. The neurons in each layer receive the output of the neurons in the layer below and send it to the neurons in the layer above [49].

3.2.2. Self-head attention

The concept of self-attention involves each element in a set being related to every other element. This is achieved through a process called "self-attention," which helps to compute a more precise representation of the set using residual connections. [46] initially introduced this idea for language understanding and machine translation tasks. This led to the development of the Transformer architecture and its various iterations, which have been widely influential in natural language processing (NLP) and computer vision.

Self-attention can be formally explained through the scaled-dot product mechanism. It involves a multiplicative attention operator that works with three sets of vectors: a set of query vectors Q , a set of key vectors K , and a set of value vectors V . Each set consists of n_k element-strong vectors created using linear projections of an identical input set of components. The key and query vectors are used to compute the similarity distribution, which is then used to calculate a weighted sum of the value vectors. This process helps to capture the relationships and dependencies between different elements in the set. [25] has further contributed to understanding self-attention and its applications.

3.2.3. Multi-head attention

The multi-head attention module in the transformer model utilizes the attention mechanism in parallel multiple times. This involves concatenating and linearly transforming the outputs of the attention mechanism. Multi-head attention allows for simultaneous self-attention across different sections of the input sequence [50], helping to capture both long-term and short-term dependencies. There are two types of attention mechanisms: soft attention and hard attention. In soft attention, weighted image features are used as input to the model instead of the raw image, enabling the model to focus on important areas and ignore less relevant ones. Soft attention uses conventional back-propagation for gradient computation and assumes that the weighted average accurately represents the focus region. On the other hand, hard attention involves sampling using the Monte Carlo approach and then averaging the results to obtain the final output. The precision of hard attention is determined by the number and quality of the samples taken [7].

The drawback of attention-based approaches is the low precision in selecting the attention area, as mentioned in some articles. Most attention-based methods choose regions of the same size and shape without considering the image contents. Determining the best number of area recommendations involves a trade-off between small and huge amounts of detail. Another issue is the single-stage structure of attention-based approaches. Since most approaches have a single encoder-decoder attention structure, they cannot generate detailed captions for the images [7]. In a typical attention-based paradigm, an adaptive attention module learns how often to attend, while a base attention model performs a single attention step for each time step. In these methods, the characteristic of the image matches one captioning word at each time step. As the output of one attention mechanism depends directly on the outcome of another, the relationship between the attended feature and the attention inquiry is not modeled [24].

The transformer model for neural machine translation highlighted multi-head attention effectiveness based on multiple scaled-dot attention heads. Both the encoder and decoder were constructed using multi-head attention. Currently, models prioritizing scaled-dot and multi-head attention over bottom-up characteristics and semantic information yield the best results for image captioning. Multi-head attention techniques outperform existing methods, making them the best practices when utilizing attention mechanisms for image captioning [31]. In Fig. 5,

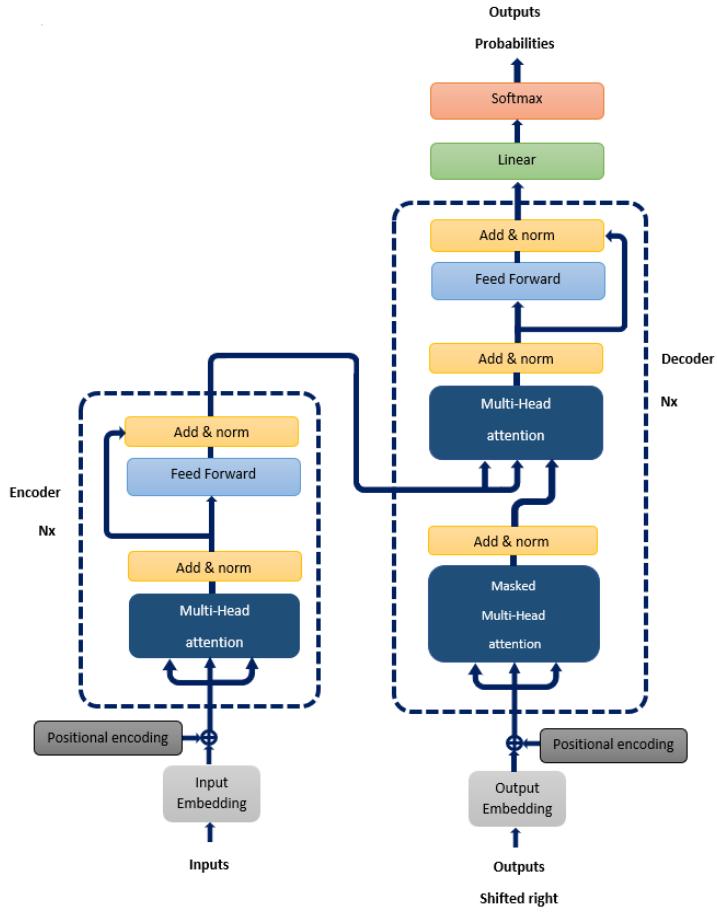


Figure 4: General architecture of an encoder-decoder transformer.

the scaled attention of the dot product is shown in the left block, where self-attention calculates the dot product of the query with all keys, which is then normalized using the SoftMax operator to obtain attention scores. These scores determine the weights, and each element becomes the weighted sum of all elements in the sequence. On the other hand, the right block represents multi-head attention, consisting of multiple self-attention blocks ($h = 8$ in the original Transformer model) to capture complex interactions between various items in the sequence.

3.2.4. Add and Norm layers

The Add and Norm layers perform two operations. The ‘add’ step controls the flow through residual connections. The second step is ‘Norm,’ which performs layer normalization. As a result, the output of this layer will follow (5).

$$\text{Add \& Norm} = \text{LayerNorm}(x + \text{Sublayer}(x)) \quad (5)$$

where x is the input of any sublayer (MHA or feedforward), and the sublayer (x) is the output.

3.2.5. Feed Forward Network

Each layer contains a fully connected point-wise feedforward network using ReLU activation for two linear transforma-

tions. The layer determines the weights used during training, which can be defined numerically as

$$\begin{aligned} FF(x) &= \text{ReLU}(xW_1 + b_1)W_2 + b_2 \\ \text{ReLU}(x) &= \max(0, x) \end{aligned} \quad (6)$$

where W_1 and W_2 are network weight matrices and b_1 and b_2 are biases.

3.2.6. Positional encoding

The transformers incorporate positional encoding to introduce the relative or absolute positions of the tokens into the model. This helps maintain the parallel execution format of the token sequence. The positional encoding values are calculated using sine and cosine functions to represent the position and training parameters. These positional encodings are combined with language features to create embeddings that are aware of the position within the sequence.

3.2.7. Linear and SoftMax layer

Like in the seq2seq models, the decoder output is transformed by a fully connected linear layer to match the vocabulary size n , representing the expected result size. The vocabulary size of a language depends on the sentence length and the size of its vocabulary. After the transformation, a SoftMax layer

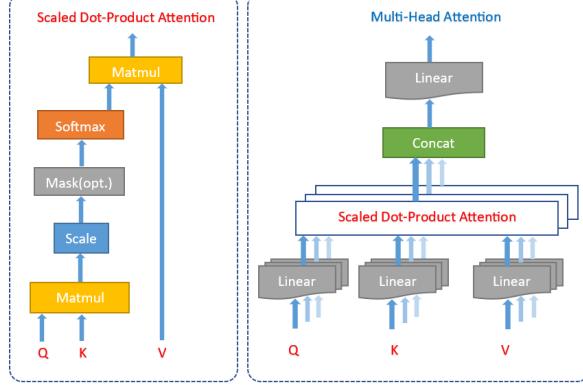


Figure 5: Attention mechanism: (left) scaled dot-product attention, (right) multi-head attention.

is applied to the resulting matrix to create a probability distribution for each word in the output phrase over the vocabulary.

3.2.8. Encoder and decoder stacks

The encoder [46] consists of a stack of identical layers $N = 6$, each containing two sub-layers. The first sub-layer is a multi-head self-attention mechanism, and the second is a simple, positionwise, fully connected feedforward network. A residual connection around the two sublayers is used, followed by layer normalization. This allows an attention vector to capture the contextual links between words in a sentence for each word. Self-attention, a specific attention mechanism used by multi-headed attention in the encoder, enables models to connect each word in the input to other words. Similarly, the decoder comprises a stack of $N = 6$ identical layers, adding a third sublayer to each encoder layer. This additional sub-layer performs multi-head attention over the output of the encoder stack. As with the encoder, residual connections are utilized around each sublayer, followed by layer normalization. Furthermore, the self-attention sub-layer in the decoder stack is modified to prevent positions from attending to preceding positions. This means that predictions for location i can only involve known outputs at positions less than i due to this masking and the offset of the output embeddings by one position. Finally, the decoder is completed by a linear layer serving as a classifier and a SoftMax to determine word probabilities.

3.2.9. Attention function

A set of key-value pairs, a query, and an output, all represented as vectors, can be linked using an attention function. The output is determined by calculating the weighted sum of the values, with each value's weight based on the compatibility of the query with its corresponding key. The attention mechanism described in [46] is called Scaled Dot Product Attention. The input consists of queries, keys of dimension d_k , and values of dimension d_v . First, the dot product of the query with all keys is computed and then divided by $\sqrt{d_k}$. Subsequently, a SoftMax function is applied to obtain the weights of the values. The attention function is continuously computed on a group of queries gathered into a matrix Q . The keys and values are or-

ganized similarly into matrices K and V . The output matrix is estimated as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \quad (7)$$

In [46], the usefulness of the linear projection of queries, keys, and values multiple times was demonstrated using distinct learned linear projections. The queries, keys, and values were projected to dimensions d_q , d_k , and d_v rather than using a single attention function with model-dimensional keys, values, and queries. The attention function was applied to these projected versions simultaneously, resulting in d_v -dimensional output values. The model could use multi-head attention to data from multiple representation subspaces at different locations. Their study used eight parallel attention layers, or heads, with the formula $d_k = d_v = d_{model}/h = 64$ applied to each. Despite using multiple heads, the total computing cost was comparable to that of single-head attention with full dimensionality due to the lower dimension of each head as shown in (8),

$$\begin{aligned} \text{MultiHead}(Q, K, V) &= \text{Concat}(\text{head}_1, \dots, \text{head}_h) W^O \\ &\text{where } \text{head} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V). \end{aligned} \quad (8)$$

3.3. Attention-based approaches

Techniques such as CNN or RNN can generate image descriptions but cannot analyze the image over time. Moreover, these approaches do not consider the spatial elements of the image that are crucial for generating image captions. Instead, attention-based techniques are gaining popularity in deep learning, as they consider the entire context when creating captions. They can dynamically focus on different elements of the input image as the output sequences are generated. These methods commonly use CNN to gather image data and then employ a language generation phase to produce words or sentences based on the output. Each language generation step focuses on the image's prominent areas until reaching the final state. Although attention-based methods aim to identify various regions of the image when generating words or phrases for image captions, the accuracy of the attention maps produced by these methods

may affect the quality of the generated captions [2]. The effectiveness of attention mechanisms in deep learning models has led researchers to emphasize their importance in image captioning [31]. Fig. 6 shows changes in attention over time as the model generates each word to reflect the relevant parts of the image [51].

People can focus on certain details while disregarding others when receiving information. This self-selection process is known as attention. The attention mechanism is an important development in generation-based models within the encoder-decoder architecture. It aims to improve the encoder-decoder model by imitating the human eye's focus on different areas in an image when generating descriptive words. The concept of attention originated from studying human vision in cognitive neurology, which led to the discovery of this higher brain function. The attention mechanism has diverse applications, including image categorization in visual images and various experiments in natural language processing, such as machine translation, abstract creation, text understanding, text classification, and visual captioning [52].

Human attention patterns and visual focus on images have inspired attention-based approaches. In these mechanisms, the model is directed to pay more attention to the most important characteristics of an image, similar to how humans do. The attention mechanism guides the model on "where to look" during the training process [7]. It is recognized that images contain a vast amount of information, but not all features need to be explained in the captioning of images. Instead, the focus should be on the most essential content. If attention is integrated into the encoder-decoder picture captioning framework, sentence creation will be influenced by hidden states computed using the attention method. This framework incorporates an attention mechanism that allows the decoding process to concentrate on specific features of the input image at each time step to generate a description of the image [1].

The human visual system inspires the mechanism of attention in image processing. Like our eyes do not take in every detail of an image at once, the attention mechanism also focuses on the key elements before moving on to the next. This approach is believed to enhance image captioning by eliminating irrelevant information. By mimicking the cognitive function of human vision, this mechanism can also reduce computational load and improve training accuracy [38].

Attention mechanisms are widely used in applications such as image captioning, machine translation, speech recognition, image synthesis, and visual question-answering models [5] [53]. Attention has been shown to connect the meaning of features, which aids in understanding how one aspect relates to another. Incorporating this into a neural network helps the model focus on the most important and relevant features while ignoring other noisy parts of the data space distribution [35].

4. Image Captioning Literature Review

This section categorizes image captioning models into Hand-Crafted Approaches for Image Captioning, Deep Learning for

Image Captioning, and Transformer-Based Image Captioning. The state-of-the-art image captioning methods are provided in Table 2. Furthermore, Table 6 offers a summary of image captioning models for different languages.

4.1. Hand-crafted approaches for image captioning

Image captioning was initially performed using traditional machine learning techniques before advancement in deep learning methods [89]. Pattern recognition systems have played an important role in solving computer vision tasks related to images [90]. Unsupervised and semantic segmentation approaches, which typically require less time and data than recent deep learning techniques, were commonly used [91]. In a study by [54], a three-stage root word-based method was proposed to generate Arabic captions for images. This involved creating image fragments using a pre-trained deep neural network on ImageNet and mapping them to a set of root words in Arabic. Furthermore, a deep belief network pre-trained by restricted Boltzmann machines was utilized to extract the most suitable words for the image [92].

4.2. Deep learning for image captioning

This section provides an overview of research papers that develop deep-learning approaches based on image captions.

4.2.1. The root words recurrent neural network and deep belief network model

In their work, [55] proposed a method for generating captions directly from images in Arabic. They utilized root-word-based recurrent neural networks and deep neural networks. The process involved extracting root words from the images, translating them into morphological inflections, and then using the dependency tree relations of these words to establish the sentence order in Arabic. They used two datasets for their study: the Flickr8k dataset, which had manually written captions in Arabic by professional Arabic translators, and a collection of 405,000 images with captions from various newspapers in Middle Eastern countries. The findings indicated that the direct one-stage generation of Arabic captions yielded better results than a two-stage process involving using English captions in the Arabic translation.

4.2.2. The convolutional neural network-gated recurrent units encoder-decoder model

To address the issues of exploding and vanishing gradients in RNN, a proposed method was introduced by [69]. The model was built upon an encoder-decoder architecture, utilizing CNN for image description and GRU (gated recurrent units) for text generation. The GRU decoder utilized an image feature vector extracted by CNN and information from the scores of phrase weights. Two methods were applied to generate the scores. The first method used the part-of-speech (PoS) technique to produce scores based on word classes, while the second method utilized a likelihood function measured by the Euclidean distance. The results indicated that the PoS approach outperformed the model.

Table 2: Overview of state-of-the-art methods in image captioning, highlighting key techniques, datasets, and performance metrics

Reference	Dataset	B1	B2	B3	B4	CIDEr	METEOR	ROUGE	SPICE
[54]	Arabic Al-Jazeera news ^m	0.348	NA						
[55]	Arabic Flickr8k ^m	0.658	0.559	0.404	0.223	NA	0.209	NA	NA
[56]	Arabic Flickr616	0.460	0.260	0.190	0.080	NA	NA	NA	NA
[57]	Arabic Flickr8k ^x	0.344	0.154	0.076	0.035	NA	NA	NA	NA
[58]	Arabic Flickr8k	0.330	0.190	0.100	0.060	NA	NA	NA	NA
[59]	Arabic Flickr8k	0.365	0.214	0.120	0.066	NA	NA	NA	NA
[60]	Arabic Flickr8k	0.443	NA	NA	0.157	NA	0.343	NA	NA
[61]	Arabic Flickr8k	0.391	0.246	0.151	0.093	0.428	0.317	0.334	NA
[62]	Arabic Flickr8k	0.391	0.251	0.140	0.083	NA	NA	NA	NA
[63]	Arabic Flickr8k	0.489	0.317	0.213	0.145	0.472	0.334	0.398	NA
[64]	Arabic Flickr8k	0.598	0.400	0.306	0.165	0.469	0.260	0.385	NA
[65]	English Flickr8k	0.579	0.383	0.245	0.160	NA	NA	NA	NA
[66]	English Flickr8k	0.589	0.335	0.263	0.148	NA	NA	NA	NA
[67]	English Flickr8k	0.674	NA	NA	0.243	0.636	0.215	0.448	NA
[68]	English Flickr8k	0.690	0.471	0.324	0.219	0.507	0.203	0.502	NA
[65]	English Flickr30k	0.573	0.369	0.240	0.157	NA	NA	NA	NA
[69]	English Flickr30k	0.695	0.463	0.341	0.232	0.486	0.302	0.451	NA
[67]	English Flickr30k	0.671	NA	NA	0.233	0.645	0.204	0.443	NA
[70]	English Flickr30k	0.677	0.494	0.354	0.251	0.531	0.204	0.467	0.145
[71]	English Flickr30k	0.647	0.456	0.320	0.224	0.467	0.197	0.449	0.136
[68]	English Flickr30k	0.689	0.468	0.319	0.220	0.428	0.191	0.487	NA
[72]	English Flickr30k	0.694	0.498	0.355	0.254	0.469	0.251	0.538	NA
[73]	English Flickr30k	0.674	0.495	0.360	0.260	0.520	0.201	0.470	NA
[74]	English Flickr30k	0.690	0.493	0.347	0.241	0.528	0.195	0.465	NA
[75]	English MS COCO	0.744	0.567	0.418	0.308	0.680	0.234	NA	NA
[51]	English MS COCO	0.718	0.504	0.357	0.250	NA	0.230	NA	NA
[76]	English MS COCO	0.748	0.577	0.428	0.314	1.061	0.265	0.553	NA
[77]	English MS COCO	0.822	0.670	0.524	0.402	1.324	0.297	0.595	NA
[78]	English MS COCO	0.828	0.681	0.536	0.414	1.360	0.301	0.604	NA
[79]	English MS COCO	0.823	NA	NA	0.398	1.319	0.297	0.598	0.230
[80]	Indonesian Flickr8k Bahasa	0.560	0.412	0.294	0.206	0.573	0.195	0.442	NA
[81]	Indonesian FEEH-ID	0.500	0.314	0.239	0.131	NA	NA	NA	NA
[82]	Indonesian Flickr8k	0.387	0.211	0.087	0.032	NA	NA	NA	NA
[83]	Indonesian Flickr8k	0.360	0.170	0.060	0.020	NA	NA	NA	NA
[84]	Myanmar Flickr8k	0.641	0.486	0.399	0.244	NA	NA	NA	NA
[85]	Myanmar corpus	0.703	0.581	0.513	0.386	NA	NA	NA	NA
[86]	Bengali BORNON	0.605	0.492	0.412	0.351	NA	0.348	NA	NA
[87]	BanglaLekhaImageCaptions	0.651	0.426	0.278	0.175	0.572	0.297	0.434	0.357
[88]	Bengali Flickr4k-Bn	0.653	0.505	0.381	0.226	NA	NA	NA	NA

^mManual extraction of Arabic dataset

Top performer in each language is in bold

^xSubset of Arabic Flickr8k (2000 images)

Top performer for each metric is underlined

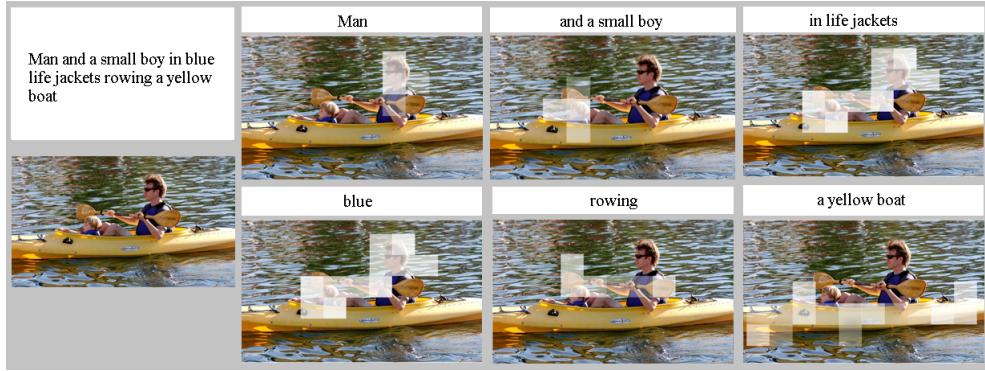


Figure 6: Attention mechanism dynamics showing how the model shifts its focus across different image regions over time, aligning its attention to generate each corresponding word in the caption.

Furthermore, [53] introduced an innovative approach that modeled the direct dependencies between caption words and image regions. This transformer-based approach could dynamically focus on various parts of the image. The proposed model included a CNN encoder to extract features from the image, and an RNN-based gated recurrent unit (GRU) was used as a decoder to simplify the model. The model was further enhanced by incorporating an attention mechanism to generate captions word by word for different image regions. This allowed the words to represent specific image regions rather than global areas, improving performance.

4.2.3. *The convolutional neural network-recurrent neural networks - long short-term memory encoder-decoder model*

The work of [93] proposed a multi-modal attention mechanism for generating news image captions. This mechanism combines visual and textual attention to generate captions from news images and text. The goal is to ensure that the caption of a news image reflects the specific event reported, making it different from a general caption. More than 98% attention was paid to the text, while the rest focused on the image.

The work of [94] introduced a text-to-picture system comprising several steps: keyword extraction, query formulation, image selection, image captioning, sentence similarity, image ranking, and image evaluation. This work identified challenges in mapping natural text to multimedia, including a lack of captions and meaningful tags for images returned from the Google search engine. To address this issue, they proposed using a deep-learning captioning model.

In the attention mechanism, determining the optimal number of regions to capture all the details in an image can be challenging. To address this issue, [6] proposed an approach that combines low- and high-level images. They used a combination of a Convolutional Neural Network (CNN) and an LSTM-based decoder to generate image captions. The visual attention mechanism is based on the history of image feature generation, and re-ranking methods were employed to measure the similarity between the generated captions and the corresponding object classes.

The work of [56] introduced a generative merge model for Arabic image captioning. This model involves the interaction

of two subnetworks to generate captions. The language model is based on RNN-LSTM to encode linguistic sequences of different lengths. At the same time, the image encoder is a fully convolutional network based on the Visual Geometry Group (VGG) that extracts image features as a fixed-length vector. A decoder model takes the fixed vectors from the previous models as input and makes the final prediction. It was suggested that this merged model could achieve excellent results for Arabic image captioning with a larger corpus.

In addition, [57] developed the Arabic Description Model (ADM) to generate full image descriptions in Arabic, compared to an earlier model based on English. The image features were obtained from CNN, and a JSON file containing image descriptions in English was translated into Arabic and fed to an LSTM network along with the CNN feature vector. The authors reported that translating recognized English captions into Arabic resulted in poor sentence structure, indicating that it is not a viable approach.

In addition, [58] developed a new Arabic image captioning dataset and evaluated two models with this dataset, demonstrating the superiority of the end-to-end model. Fig. 7 illustrates the proposed model employing a sequence-to-sequence encoder-decoder framework for image captioning. This involves encoding the input image into a feature vector using CNN and decoding that feature vector into an Arabic sentence using RNN.

An automatic model that converts standard Arabic children's stories into representative images that support the meaning of the words was proposed by [95]. The method consists of seven steps: Keyword extraction, query formulation, image selection, captioning, sentence similarity, image ranking, and image evaluation. Teachers or parents can use this system to help children review the materials they have studied in school.

In a separate study, [4] presented recent work on Arabic image captioning. Their research introduced an architecture-based encoder-decoder that outperforms classical methods using the standard Neural Machine Translation (NMT) approach. This approach used a CNN as an encoder to extract visual information from the input image. At the same time, an LSTM acts as a decoder, producing a probability distribution over possible next steps to generate the caption. The proposed active learning framework involved human annotators to refine the automatic

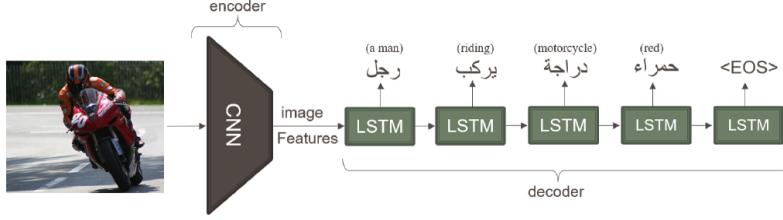


Figure 7: Sequence-to-sequence encoder-decoder framework for Arabic language image captioning [58].

translation produced by the model.

The automatic captioning of images in Indonesian was developed by [81]. The model comprises three components: an image extractor that generates feature vectors using CNN, a sequence processor that encodes linguistic sequences based on LSTM using the output from the previous step, and a decoder that predicts the caption of a new image based on the vector image features and vocabulary input. The test set showed promising results.

The work of [3] introduced a new Vietnamese image captioning method. This method comprises an image captioning model, an English-Vietnamese translation model, and an unknown word processing model. The image feature extractor utilizes CNN, and the translation model comprises an encoder-decoder, RNN. Additionally, the model provides an unknown word processing module to address the problem of unknown words in Vietnamese translation.

Myanmar's [84] proposed an image captioning method combining two parts: CNN for image feature extraction and LSTM for text generation. New datasets were built based on the Flickr8k dataset. The new datasets used 3,000 images from the Flickr8k dataset, each with five annotated Myanmar captions. This approach reduced the manual captioning time by translating the sentences. The generated text was evaluated using BLEU, and satisfactory results were obtained.

A new model for Bengali image captioning was proposed by [88]. The model utilized two-word embedding techniques and consisted of a two-part encoder and decoder. The encoder comprised a convolutional neural network, while the decoder included BiLSTM and BiGRU. The process involved extracting the image features and concatenating the output word vectors, which were then passed to the decoder after aligning the dimension between the word vector and the image features. The decoder utilized the concatenated output to generate the next word in the sequence with the highest probability. The Flickr8k dataset was used for testing, with five captions for each image translated into Bengali using Google Translator.

Bengali Image Captioning (BIC) was also presented in [87]. The model consisted of an image feature encoder, a word sequence encoder, and a caption generator. The model was tested on the BanglaLekhaImageCaptions dataset, which contains 9,154 images, each with two captions generated by two native Bengali speakers.

4.2.4. A summary of deep learning methods

In most methods, the image is first fed into a CNN to generate image features, which will then be used as input for the language processing component. The convolution layer reduces the image into features by using information from nearby pixels. It then employs prediction layers to forecast the target values. This is achieved by creating a dot product using multiple convolution filters, or kernels, which scan the image and extract unique aspects of the image. The max pooling layer helps to reduce the spatial size of the convolved features and prevents overfitting by providing an abstract representation of the convolved features. Although there are many different activation functions, ReLU is the most commonly used one in various types of neural networks due to its ease of training and superior performance due to its linear behavior [96].

In a CNN network, the higher layers are believed to capture high-level semantic information. As a result, the output of the fully connected layer can represent the image's global information. However, since this output lacks spatial information, the output of the last convolutional layer is often utilized. This is because the expanded receptive field of the higher layer in CNN corresponds to a region of the original image, where each point on the spatial feature map corresponds to a region of the original image [97]. The architecture of the CNN model is shown in Fig. 8.

The general architecture of image captioning models that use the encoder-decoder framework is depicted in Fig. 9. The encoder comprises a CNN for extracting image representations, while the decoder incorporates an LSTM for generating image captions. CNNs are a type of feedforward artificial neural network that is adept at processing visual data. A typical CNN consists of an input, an output, and multiple hidden layers. The hidden layers of a CNN typically include convolutional, pooling, fully connected, and normalization layers. On the other hand, text generation is handled by an essential deep learning model capable of learning long-term dependencies, the LSTM. An LSTM consists of a cell, an input gate, an output gate, and a forget gate as its internal components. Using simple learned gating functions, the internal units of an LSTM utilize nonlinear mechanisms to enhance hidden states, allowing them to propagate unchanged, be updated, or be reset.

4.3. Transformers-Based Approaches for Image Captioning

This section reviews research papers that focus on developing transformers that generate captions.

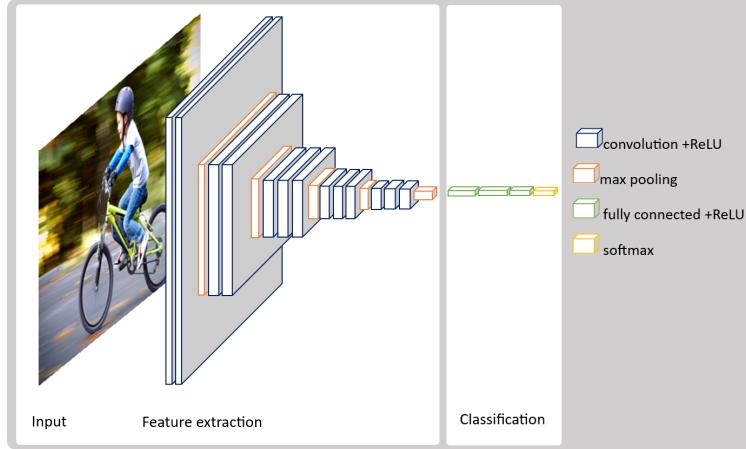


Figure 8: A typical architecture of the CNN model.

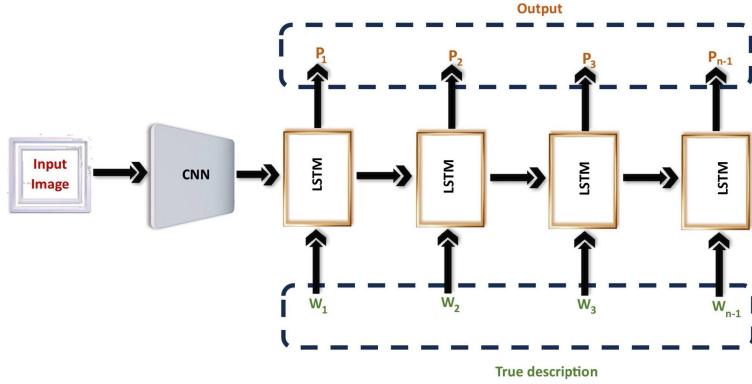


Figure 9: A general architecture of the CNN-LSTM encoder-decoder model for image captioning.

4.3.1. CNN-transformer encoder-decoder model

The study by [5] introduced a multi-transformer (MT) for image captioning. This MT model can understand three types of relations: word-to-word, object-to-object, and word-to-object. The transformer mechanism consists of an image encoder and a text decoder. The image encoder has two parts: an aligned multiview encoder and an aligned multi-view decoder. The caption decoder takes the output from the encoder and generates a caption using word embedding and one layer of LSTM.

In a different approach, [50] utilized a faster region-CNN (R-CNN) to extract visual features for a given image. These features are then inputted into the transformer encoder, allowing the transformer to effectively capture object information by overcoming interference from non-critical objects. The attention matrix computed from the transformer encoder is passed into the attention gate, where the attention weight values below the gate threshold are truncated. The decreasing threshold leads to the preservation of more non-zero values, expanding the attention scope of the self-attention module from local items to all objects as the network layer expands.

In traditional practices, normalization has been applied outside of self-attention. However, a study by [98] introduced

a novel normalization method and demonstrated its feasibility and advantages for hidden activation within self-attention. They proposed a geometry-aware self-attention (GSA) class that extends self-attention to explicitly consider relative geometry relations between objects in an image for feature extraction.

The work in [99] employed a dual-modal transformer to capture intra- and inter-model interactions within an attention block. They concatenated two embeddings, one based on the image's objects and the other using an Inception-V3 model, to create the final image-based embedding. The study showed that this model outperformed established models such as encoder-decoder and attention models.

State-of-the-art techniques directly encode identified object regions and utilize region features. However, this approach presents challenges related to object relationships and the potential for incorrect item detection. Significant computational power is also required to compute region features, particularly when using high-performing CNN-based detectors like Faster R-CNN. The study by [100] addressed these issues by replacing a CNN backbone with a transformer to overcome the drawbacks of CNN-based detectors and reduce computational costs for extracting initial features from input images.

4.3.2. LSTM-transformer encoder-decoder model

In their work, [47] introduced a new transformer-based model that considers the relationships between different features within an image. This model considers three types of spatial relationships in the image regions: a query region can be a parent, neighbor, or child. The model uses spatially adjacent matrices to combine the output of parallel subtransformer layers. The decoder includes an LSTM layer and an implicit transformer layer, which work in parallel to decode different image regions.

Two new geometry-aware architectures were separately created for the encoder and decoder to represent geometry better [101]. This captioning model helps us understand the locations of target objects and the objects the model is currently looking at. The proposed model includes an improved encoder and may provide information on an object's relative geometry. Furthermore, it fully leverages geometry relations to enhance object representations.

Remote sensing image captioning (RSI) aims to generate descriptions of the information contained in RSIs automatically. The multiscale information of RSIs encompasses the properties and complex relationships of items of various sizes. The study by [102] developed a new model based on the encoder-decoder framework. In this model, ResNet50 serves as the encoder to extract multi-scale information. At the same time, a multi-layer aggregated transformer (MLAT) is employed in the decoder to construct sentences using the extracted data effectively. Additionally, LSTM aggregates features from multiple transformer encoding levels to enhance feature representations.

4.3.3. Transformer-based model

In their work, [103] proposed a novel transformer-based approach to address the limitations of recurrent neural networks (RNN). They introduced an attention mechanism that combines visual and semantic attention to handle complex relationships. Since not every word has a corresponding visual signal, taking into account semantic information is crucial. The proposed method includes a control mechanism for the forward propagation of multi-model information. The model utilizes a dual-way transformer encoder to investigate inter- and intra-relationships between visual and semantic attributes. The decoder's output is passed to a classifier to predict the next word.

In [77], a transformer-based model was introduced for image captioning. The approach involved using a mask operation to automatically evaluate the impact of the features of the image region and using the results as supervised information to guide attention alignment. The basic version of the transformer was utilized in this study. The researchers investigated the relationship between attention weights and feature importance metrics in image captioning to comprehensively analyze whether current attention mechanisms can focus on crucial and effective image regions. This work serves as a valuable reference for self-supervised learning.

The research by [104] introduced an attention-based approach. The model is designed to capture dependencies within image areas and between image regions and external states. Using the self-attention method, the captioning model can identify

the most relevant regions at each time step. The researchers explored a sequential transformer framework based on the original transformer structure, combining the decoder with outside-in attention and RNN. The study revealed that the transformer's self-attention allows for the simultaneous direct calculation of relationships between internal areas, thus avoiding recurring attention issues.

The LATGeO framework, based on transformer technology, was introduced in [105] to generate captions for images. It incorporates multi-level geometrically coherent and visual recommendations to establish relationships between objects based on their localized ratios. A new label-attention module (LAM) was developed to connect the visual and linguistic aspects to extend the traditional transformer. In this proposed approach, object labels are included as input data at each decoder layer to assist in constructing captions.

In the context of image captioning, [106] proposed a Multi-Gate Attention (MGA) block within a pre-layer norm transformer architecture. This architecture modifies the standard self-attention mechanism by incorporating multiple gate mechanisms, thus enhancing its capabilities. The pre-layer norm transformer architecture differs from the original transformer architecture in that the layer normalization is placed before the self-attention module, and the feedforward layer and subsequent layers are eliminated. This simplification aims to increase the model's efficiency for image captioning.

The Transformer architecture, which was recently announced, utilizes self-attention to enhance the performance of sequence-analysis tasks. This has led to exploring transformers in [107]. The experimental validation was conducted using the caption dataset from the University of California (UC)-Merced. The proposed technique can potentially generate helpful textual descriptions for remote-sensing images.

In transformer-based image captioning, three-parameter reduction techniques were utilized [108]. Firstly, the size of the embedding matrices was significantly reduced by using radix encoding, allowing for a larger vocabulary without increasing the model size. Secondly, cross-layer parameter sharing was employed to break the tight correlation between model depth and size, allowing additional layers to be added without increasing the parameter count and vice versa. Finally, attention parameter sharing was used to reduce the parameter count of the multi-head attention module and improve overall parameter efficiency.

To effectively capture complex interactions within and between input features in images, a Modular Co-Attention Transformer Layer (M-CATL) was proposed by [109]. This layer aims to extract specific image characteristics. Furthermore, a Deep Modular Co-Attention Transformer Block (DM-CATB) was developed and integrated into the encoder part of the model based on M-CATL. To fully capture spatial and positional information of image features and improve feature characterization, a Deep Modular Co-Attention Transformer Network (DM-CATN) was introduced.

Local visual modeling with grid features is crucial to generating comprehensive and detailed image captions. In their work, [67] proposed a locality-sensitive transformer network

(LSTNet) to facilitate local object recognition during captioning. They also employed layer-specific fusion (LSF) for cross-layer semantic complement, combining information from multiple encoder layers. The experimental results demonstrated that LSTNet’s local visual modeling outperformed many state-of-the-art captioning models.

The study by [110] introduced an improved architecture for image captioning by incorporating a unique memory mechanism into a Transformer-based framework, addressing the challenges of maintaining long-range relationships and contextual coherence in traditional image captioning algorithms. The authors proposed the Meshed-Memory Transformer (MMT), which integrates a memory module to improve the model’s capacity to retain and utilize data in both temporal and spatial dimensions. This memory-enhancing mechanism and a typical Transformer model helped the MMT system capture complex links between generated text and visual elements, leading to more detailed and cohesive captions. The research demonstrated that MMT significantly improved captioning performance in various benchmark datasets.

Additionally, a technique called the full memory transform was described in the work by [49]. This technique aims to enhance the efficiency of language decoding and image encoding. The Full-Layer Normalization Symmetric Structure for Image Encoding was suggested, embedding Layer Normalization symmetrically on both sides of the self-attention network (SAN) and feedforward network for robust training and higher model generalization performance. Furthermore, the Memory Attention Network was introduced to extend the conventional attention mechanism, directing the model to concentrate on words that require attention, thus improving the language decoding step.

4.3.4. Summary of transformer-based models

Transformers generate the words of the caption all at once, while model-based RNN still produces the caption word by word (see Fig. 10). In the model on the left, which is a CNN-RNN-based model, the caption words are produced one by one [51]. On the other hand, the model on the right demonstrates the transformer’s ability to generate a full text with all words simultaneously [111].

Fig. 11 shows a general architecture of the transformer model for image captioning. It includes a feature extraction model, typically a CNN, and a transformer for text generation. The transformer comprises an image encoder to learn self-attended visual features and a caption decoder to generate the caption from the attended visual features.

4.4. Attention-based approaches for image captioning

The study by [112] introduced a new approach to address the computational limitations of traditional attention mechanisms in image captioning. The authors presented X-Linear Attention Networks (X-LAN), which combine a linear attention module for improved computational efficiency and reduced complexity with a non-linear module capturing more detailed interactions and dependencies within the image. The study

demonstrated that X-LAN produces significant performance improvements in benchmark datasets compared to existing methods, offering a more scalable and effective solution to generate detailed and contextually accurate image descriptions. By enhancing both efficiency and accuracy, X-LAN advanced the capabilities of image captioning systems.

A separate study by [113] improved the attention mechanisms used in image captioning. They proposed a novel framework called Attention on Attention (AoA), which enhances existing models by introducing a secondary attention mechanism that acts on the primary attention outputs. This secondary process reassesses and recalibrates the original attention weights, considering the generated words’ context and visual elements’ context.

In their work, [114] proposed a new method called Reflective Decoding Network (RDN) to enhance image captioning systems. Unlike traditional models, which often employ a single-stage decoding process that may not fully utilize the context and finer details of the visual input, RDN involves a two-step decoding process. In the first stage, a reflective mechanism generates an initial caption, followed by a second stage to refine it. This reflective decoding process employs a self-attention-based approach to review and modify the original caption, considering the visual elements and previously generated words. This iterative refinement results in improved captioning output from the model.

Developing non-visual words such as "to" and "itself" does not require much visual information. Therefore, using image features as key-value pairs for cross-attention to create captions for images is unsuitable. In the Task-Adaptive Attention model proposed in [115], task-adaptive vectors were included to learn nonvisual signals that can help address this issue in image captioning. The comprehensive Transformer model with Task-Adaptive Attention integrates the suggested task-adaptive attention module into a standard transformer-based encoder-decoder architecture.

In a study by [116], a novel image captioning technique called Dynamic Attention Prior (DY-APR) was introduced. This approach combines attention distribution before the local linguistic context for dynamic attention aggregation. The researchers proposed a method for dynamically aggregating the Attention Distribution Prior (ADP) and the current layer’s attention score to provide more precise attention guidance. They also presented a learning technique to gradually transition input tokens from a fully static representation based on word embedding to a mixed scheme incorporating both the input tokens and the linguistic context.

Existing image captioning methods focus primarily on the visual attention mechanism, often resulting in incomplete and inaccurate model-generated sentences. In addition, errors in extracting visual features can lead to incorrectly generated captions. The work of [117] addressed this gap by proposing a combination attention module consisting of two modules: visual attention and keyword attention. The evaluations demonstrated that this strategy yielded better results.

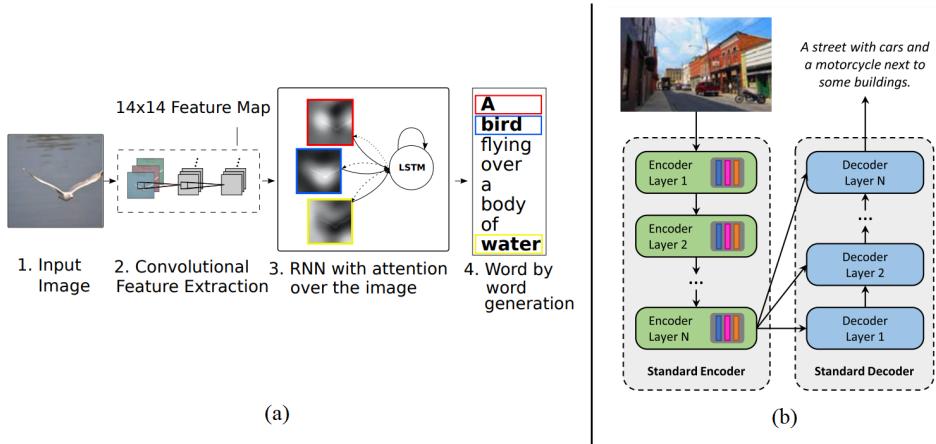


Figure 10: Image captioning models: (a) CNN-RNN based model [51], (b) transformer-based model [111].

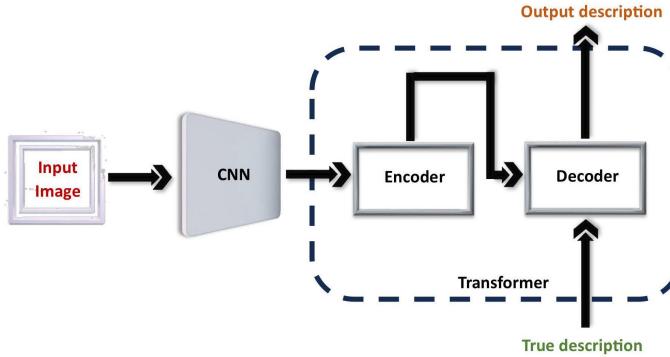


Figure 11: General architecture of CNN-transformer model for image captioning.

4.4.1. Soft and hard attention

The first attentive deep paradigm for image captioning was Show, Attend, and Tell [118]. In this model, the decoder used an LSTM for language modeling, and the feature extractor was a CNN. Specifically, the VGG model was pre-trained on ImageNet. Show, Attend, and Tell was quite similar to other CNN-LSTM encoder-decoder architectures for captioning videos, except that it utilized two attention mechanism variants: soft and hard attention on the spatial convolutional features to generate a set of attended features for the LSTM decoder, acting as a language model.

In Show, Attend, and Tell, attention involves a set of attended visual features (z) generated from an attention function (fatt), which can be soft deterministic or hard and stochastic. In soft attention, the input to the LSTM comprises weighted image characteristics that take attention into account rather than the image x . Soft attention reduces the weight of irrelevant places with low attention, which helps to focus on relevant areas.

In soft attention, areas of high focus retain their original values, while areas of low focus approach 0. This is achieved by assigning a weight, a_i , to each x_i input to the LSTM. The sum of all weights, a_i , is 1, representing the likelihood of focusing on x_i . On the other hand, hard attention uses a stochastic sampling model by selecting x_i as input to the LSTM, with a_i serving as

a sampling probability rather than a weighted average.

The Monte Carlo approach is used in hard attention to accurately calculate the gradient descent during backpropagation, while soft attention uses the standard backpropagation method [118]. This allows the model to concentrate its computation on specific salient regions while generating captions, using soft and hard attention to understand the concept of attention in image annotation.

Soft attention can be trained through standard backpropagation by applying weights to the annotated vector of picture features when the feature is salient. In contrast, stochastic hard attention can be trained by maximizing the lower bound variation [35].

It is important to note that the spatial information extracted from the two-dimensional image is crucial for both soft- and hard-attention mechanisms. The extracted annotation vector contains features of each color channel in a 3-dimensional spatial feature vector since the images are represented using three color channels (red, blue, and green). Once the set of spatial features attended is determined, they are ready for use [31]. Soft and hard attention can be seen in Fig.12. In this illustrative figure, you can observe how soft attention demonstrates the relative importance of each part of the image to the other parts. In contrast, hard attention separates specific parts of the image

and considers only these parts when generating the next word in the caption, disregarding the rest [51].

4.4.2. Bottom-up and top-down approaches

Employing saliency is based on how our brain processes visual information. It combines a bottom-up flow of visual inputs with a top-down reasoning process. The top-down approach involves predicting incoming information from the environment using our prior knowledge and logical bias. In contrast, the bottom-up approach involves visual signals that correct the prior predictions. Additive attention can be approached as a top-down system, where the language model observes a feature grid independent of the image content and predicts the subsequent word. The bottom-up path is defined by an object detector responsible for identifying image regions. Then, a top-down process learns to weigh each region for each word prediction. This approach is connected with the concept of additive attention [25].

Scientists have focused on top-down and bottom-up attention theories, and recent research has shown that top-down attention mechanisms are still preferable. The top-down model starts with an image as input and converts it to words [36]. All methods that utilize bottom-up attention perform better because bottom-up attention focuses on visual attention at the object level. However, an important question arises: In some natural settings, is it necessary for the model to pay attention to areas in the image that do not contain recognizable objects but instead include natural elements such as mountains, trees, skies, etc.? On the other hand, using object detectors for bottom-up feature extraction has drawbacks, as they may not be able to focus on important areas for captions in unfamiliar domains. As a result, additional knowledge from various domains may be necessary for natural settings, and object detectors trained in more specialized tasks can provide this kind of knowledge [31].

Attention-based encoder-decoder models are known for their sequential information processing but are criticized for lacking global modeling skills. To overcome this limitation, a reviewer module has been developed to conduct review stages on the encoder's hidden states and generate a thought vector at each step. The attention mechanism achieves this by assigning weights to the hidden states. The thought vectors capture global aspects of the input and effectively review and learn the information encoded by the encoder. The decoder uses these thought vectors to predict the next word [1]. Additionally, incorporating visual attention allows for a multimodel coverage mechanism [93]. This visual attention mechanism uses features derived from a convolutional neural network layer, where each feature represents an abstraction of a region in the image and provides a weighting for each geographical region. A higher weight indicates a more important image region [6]. It is worth mentioning that the described attention method falls between the encoder and the decoder.

Figure 13 illustrates an example of bottom-up and top-down approaches. A set of salient image regions is identified in the bottom-up approach, and a pooled convolutional feature vector, like Faster R-CNN, an exemplary bottom-up attention mechanism, describes each region. The top-down approach, on the

other hand, utilizes the task-specific context to determine an attention distribution over the visual regions. The weighted average of the image features in all regions is then utilized to compute the attended feature vector. The study by [119] proposed a method that presented bottom-up and top-down approaches.

4.4.3. Summary of attention-based models

Figure 14 illustrates the general architecture of the attention model. This innovation has greatly enhanced image captioning, allowing the algorithm to focus on important image aspects and ignore redundant content. This model implements attention as a weighted sum of encoder outputs. A CNN first processes the image within the encoder-decoder framework, resulting in feature maps. Subsequently, the attention module assigns a weight to each image pixel based on the feature maps and a hidden state. These weights enable the decoder to generate words for the output text while concentrating on the most pertinent parts of the image.

4.5. Graph-based representation for image captioning

The study by [120] emphasized the importance of visual relationships among objects, advancing the field of image captioning. Traditional image captioning models typically focus on object detection and identification, generating descriptive text based solely on these aspects. However, such approaches often neglect the intricate connections and interactions between objects that can greatly enhance the depth of the captions. The authors introduced a new approach integrating a visual relationship module into the captioning architecture to address this limitation. This module analyzes and encodes the interactions between elements in an image using a graph-based representation. This enables the model to understand better and express the spatial and functional relationships between items, resulting in more detailed and contextually accurate captions. The research offered a comprehensive analysis of their methodology, demonstrating significant improvements in relevance and caption quality compared to existing approaches. The authors expanded the boundaries of current image captioning systems by showcasing through extensive experiments that incorporating visual relationships enhanced the descriptive power of the captions and improved the model's ability to generate coherent and contextually appropriate descriptions.

4.6. Comparative analysis of state-of-the-art methods for image captioning

This section evaluates the effectiveness of various state-of-the-art methods for Image Captioning, as presented in Table 2. The table includes numerous methods used on different datasets, including Arabic datasets such as Flickr8k and Flickr30k and English datasets like Flickr8k, Flickr30k, and MSCOCO, along with other datasets from various languages. Performance was measured using several metrics, including CIDEr, METEOR, ROUGE-L, SPICE, and BLEU scores at four levels (BLEU-1 through BLEU-4). Upon a thorough examination, it is evident that the technique [55] achieved a high BLEU-1 score of 0.658 when applied to the Flickr8k Arabic dataset using manual extraction. Compared to non-English datasets, methods applied

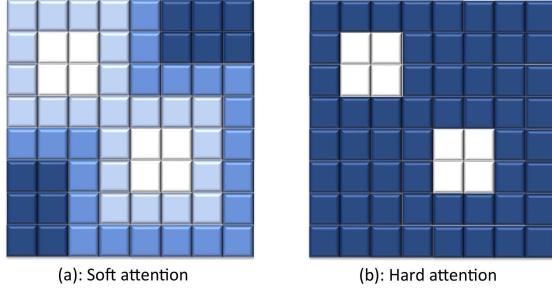


Figure 12: Examples of soft and hard attention mechanisms. (a) Soft attention assigns varying importance to different parts of the image, influencing the entire caption. In contrast, (b) hard attention focuses on specific regions, selectively considering parts of the image while ignoring others.

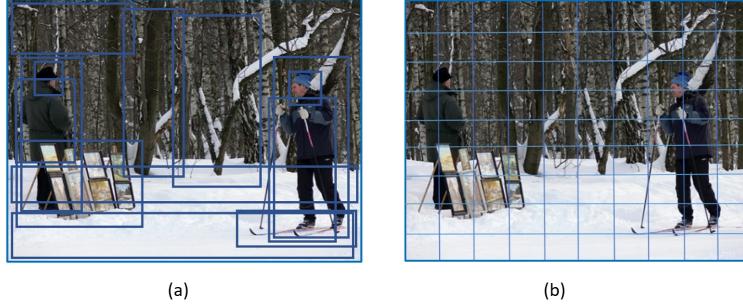


Figure 13: Illustrations of bottom-up and top-down attention approaches. (a) Bottom-up attention, where focus is determined at the level of objects and other salient regions of the image, and (b) top-down attention, where features correspond to a uniform grid of equally sized image regions.

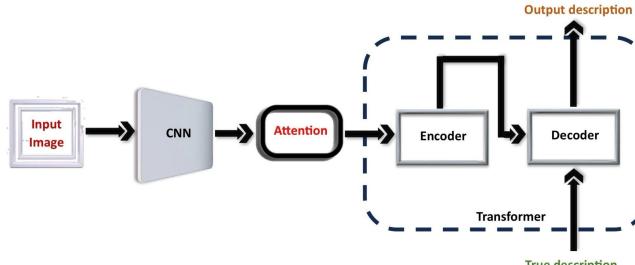


Figure 14: General architecture of an attention-based model for image captioning, illustrating the integration of image features with sequential attention mechanisms to generate descriptive captions.

to English-based datasets like Flickr8k and Flickr30k generally yield better scores across most criteria.

In the English Flickr8k dataset, the method described in [68] achieved a high BLEU-1 score of 0.690, giving a strong performance in generating relevant captions. However, with the Arabic Flickr8k dataset, the method in [64] only achieved a BLEU-1 score of 0.598, revealing the challenges of adapting methods to different languages and contexts. Additionally, when the method described in [88] was applied to the Bengali Flickr8k dataset, it produced lower scores, with many values marked as 'NA,' suggesting that all metrics did not evaluate the approach.

Further analysis shows that methods evaluated using the English MS COCO dataset, including the method [78], generally achieve high scores on various metrics, such as a CIDEr score of 1.360 and a BLEU-4 score of 0.414. This indicates that the

MS COCO dataset, a widely used and comprehensive dataset, provides a reliable standard for evaluating image captioning methods.

Furthermore, with the MS COCO dataset, techniques such as [79] and [77] demonstrate strong performance, scoring highly in the BLEU and CIDEr metrics, showcasing their effectiveness in generating diverse and accurate captions. However, achieving high performance on non-English datasets such as Bengali BORNON and Indonesian FEEH-ID is more challenging, underscoring the need for further research and development in multilingual and culturally diverse image captioning techniques. This variation underscores the importance of creating more inclusive datasets and methods that perform well in linguistic and cultural contexts.

The variation in results between metrics for the same approach suggests that no single measure can fully evaluate the

effectiveness of image captioning methods. For example, a method might have a high BLEU score but a low CIDEr or SPICE score, indicating different strengths and weaknesses. This highlights the need for robust and adaptable methods for diverse datasets and languages. Comparative analysis underscores the importance of adapting image captioning methods to specific contexts. Although some methods perform well on certain datasets, they may not perform as well on others, emphasizing the importance of continued development and adaptation in image captioning.

5. Datasets

This section introduces the commonly used datasets in image captioning. Table 3 illustrates the datasets’ details.

5.1. English datasets

5.1.1. The Flickr8K dataset

The Flickr8K dataset, developed by [121], was publicly released in 2013. It comprises 8,000 images sourced from the Flickr image-sharing platform. Compared to MS COCO, Flickr8K is relatively tiny and primarily contains photographs of humans and animals. The image descriptions were manually annotated using Amazon Mechanical Turk, with each image paired with five descriptive sentences, ensuring linguistic diversity in the captions.

5.1.2. The Flickr30k dataset

The Flickr30K dataset [122] is an expanded version of Flickr8K, containing 31,783 captioned images. Each image is accompanied by five descriptive sentences, providing a diverse linguistic representation. The dataset primarily consists of photographs depicting people engaged in everyday activities and events, making it a valuable resource for training and evaluating image captioning models.

5.1.3. The Microsoft COCO datasets

The Microsoft COCO (MS COCO) dataset [123] is a large-scale benchmark widely used in image recognition, object detection, semantic segmentation, and image captioning. Each image is manually annotated via Amazon Mechanical Turk and includes objects from over 100 categories, representing real-world scenes with natural backgrounds. The dataset contains 82,783 training images, 40,504 validation images, and 40,775 test images with undisclosed labels. Each image is paired with five descriptive captions, making MS COCO a key resource for evaluating image captioning models.

5.2. Arabic datasets

The model proposed by [56] was trained and tested using images from the MS COCO and Flickr8K datasets. The MS COCO dataset contains over 330,000 images and 2.5 million captions, covering 80 object categories. The CrowdFlower crowdsourcing service was used to generate Arabic captions, resulting in 5,358 captions for 1,166 images from the training set, with an average of 4.6 captions per image.

The Flickr8K dataset, which consists of 8,000 images, initially includes five English captions per image. The first 2,261 images from its training set were selected for this study, and a professional translator created 750 Arabic captions. The remaining images were translated into Arabic using Google Translate, followed by manual verification by native speakers. In total, 3,427 images (from both MS COCO and Flickr8K) were used, with a vocabulary size of 9,854 words, and the longest caption containing 27 words. For experiments, the dataset was split into 2,400 training images (70%), 411 development images (12%), and 616 test images (18%).

Two test scenarios were proposed in [57]: (1) *Machine translation approach* – English image descriptions were translated into Arabic using Google Translate, often leading to grammatical errors and poorly structured sentences. (2) *LSTM-based Arabic generation* – Instead of relying on translation, an LSTM-based Arabic language model was trained to generate more natural and grammatically accurate captions.

To evaluate the model’s performance, three distinct test sets were used: (1) *Simulated trained images* – The model was tested using trained images from Flickr8K. (2) *Unseen test images* – The model was evaluated using images from the Flickr8K test set that were not seen during training. (3) *Tishreen University dataset* – A new test dataset was created using images from Tishreen University, further improving the reliability of the experiments.

The work by [58] introduced the Arabic Flickr8K dataset. To create this dataset, the original English Flickr8K dataset was translated into Arabic using a two-phase process: First, the Google Translate API produced an initial Arabic translation. Next, qualified Arabic translators carefully reviewed and refined the translations. After verification, the top three translations per image were selected from an initial pool of five. The final Arabic Flickr8K dataset consisted of 6,000 training images, 1,000 validation images, and 1,000 test images, each with three unique captions.

In a separate study, [4] developed the ArabicFlickr1K dataset using an active learning-based framework to translate an existing dataset. The final ArabicFlickr1K dataset contains 1,095 images, with three to five Arabic captions per image. This dataset was designed to support Arabic image captioning models, offering a diverse and linguistically rich dataset for improved training and evaluation.

5.3. Other languages datasets

The Vietnamese image captioning model proposed by [3] was evaluated using the UIT-ViIC dataset, which was carefully curated to ensure consistent and accurate captions. The annotation process was conducted by five native Vietnamese speakers (aged 22–25) trained in sports-related vocabulary before starting. The dataset consists of 3,850 sports-related images sourced from the 2017 Microsoft COCO edition, with each image accompanied by five Vietnamese captions, totaling 19,250 captions. To minimize inconsistencies in interpretation, strict annotation guidelines were established, inspired by the MS COCO dataset. These included: First, a minimum of ten Vietnamese

words per caption. Second, captions should describe only visible objects and activities, excluding personal opinions, proper names, and numbers. Last, sentences should be written in continuous tense, with familiar English terms (e.g., "tennis") allowed for clarity.

The Indonesian dataset used in the study by [81] is FEEH-ID, which contains 8,099 images, each paired with five captions in Indonesian. The images were sourced from Flickr. The first 6,000 images from the training set were selected for their experiments. The captions were generated using a combination of Google Translate and manual translation by a professional English-Indonesian translator. The total vocabulary size of the dataset varied depending on the frequency of objects appearing in the images.

The Myanmar image captions corpus, developed by [84], was built using a subset of the Flickr8K dataset, which initially contains 8,092 images, each with five English captions. Due to time constraints, 3,000 images were selected, and five Myanmar-language captions were created for each image, totaling 15,000 captions. The dataset was constructed using two approaches: First, English captions were translated into Myanmar using an attention-based neural machine translation model, achieving a multi-BLEU score rate 13.93. Second, native speakers directly described the images in Myanmar, resulting in a vocabulary size of 3,138 words, with the longest caption containing 21 words. The dataset was divided into 2,500 images for training, 300 for validation, and 200 for testing.

The study by [86] utilized three key datasets to generate Bengali captions for images. Together, these three Bengali datasets provide a comprehensive and diverse collection of images and captions, enabling more accurate and contextually rich Bengali caption generation. The datasets are:

- a) Flickr8K-BN dataset: Contains 8,091 images, each with five Bengali captions. It covers a wide range of topics, including people, landscapes, animals, and everyday objects. The captions were originally in English and later translated into Bengali.
- b) BanglaLekha dataset: Comprises 9,154 images, each with two Bengali captions. It focuses on themes such as animals, birds, food, trees, and buildings. It features a smaller vocabulary size than other datasets due to fewer captions per image.
- c) Bornon dataset: Contains 4,100 images, each with five Bengali captions, totaling 20,500 captions. It covers diverse topics, including animals, people, food, weather, and vehicles. The Images were sourced from a personal photography club, and 17 native Bengali speakers annotated captions.

6. Evaluation Metrics

Measuring the accuracy of a generated text in describing an image is done more effectively through direct human judgments. However, expanding human evaluation is difficult due to the high amount of nonreusable human effort required. The following subsections introduce the commonly used evaluation

Table 3: Publicly available datasets for image captioning, detailing dataset names, sizes, and number of captions.

Datasets	Train	Validate	Test	Captions
Flickr8k	6,000	1,000	1,000	5
Flickr30k	29,783	1,000	1,000	5
MS COCO	113,287	5,000	5,000	5

metrics in image captioning. Table 4 provides definitions of key symbols used in the evaluation metrics. These definitions help understand the mathematical formulations behind BLEU, METEOR, CIDEr, ROUGE, and SPICE.

6.1. Bilingual Evaluation Understudy BLEU

Bilingual Evaluation Understudy (BLEU) is a metric used to evaluate the quality of machine-generated text. Assess individual text segments by comparing them to reference texts. The BLEU score varies depending on the number of reference translations and the length of the text produced. Generally, short-generated texts have higher BLEU scores ranging from 0 to 1. BLEU-1 uses unigram comparisons between candidate and reference sentences, while bigram comparisons are used for BLEU-2. An empirical maximum order of four optimizes correlation with human judgments. Unigram scores determine the adequacy of the BLEU metrics, while higher n-gram scores determine fluency [124]. The BLEU formula is defined as

$$BLEU = BP \times \exp \left(\sum_{n=1}^N w_n \log (p_n) \right). \quad (9)$$

The shortness penalty (BP) allows us to choose the candidate translation most similar to the reference translation in terms of length, word choice, and word order. It is calculated using an exponential decay given as

$$BP = \begin{cases} 1 & m_c > m_r \\ e^{(1-m_r/m_c)} & m_c \leq m_r \end{cases} \quad (10)$$

The sum of the counts of the clipped n gram of candidate sentences in corpus CC_n is divided by the total number of candidate n -grams. C_n is used to calculate the modified precision for each n -gram. It enables us to determine the sufficiency and fluency of the candidate translation relative to the reference translation as

$$p_n = \frac{\sum_{C \in \{\text{Candidates}\}} \sum_{n\text{-gram} \in C} CC_N}{\sum_{C' \in \{\text{Candidates}\}} \sum_{n\text{-gram} \in C'} C_N}. \quad (11)$$

6.2. The Recall Oriented Understudy for Gisting Evaluation ROUGE

The Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [125] is a set of metrics used to assess text summaries. It compares word sequences and word pairs with a reference database of human-written summaries in a given summary. The metric uses the longest common subsequence between a candidate

Table 4: Definitions of symbols in evaluation metrics.

Symbol	Definition	Used In
n -gram	Sequence of n consecutive words in a sentence	BLEU, ROUGE
P_n	Precision of n -gram matches between generated and reference captions	BLEU
BP	Brevity penalty (penalizes overly short captions)	BLEU
R	Recall – fraction of reference words covered by the candidate caption	METEOR, ROUGE
P	Precision – fraction of candidate words appearing in the reference caption	METEOR, ROUGE
F_1	Harmonic mean of precision and recall: $F_1 = \frac{2PR}{P+R}$	METEOR, ROUGE
weight(n)	Weight assigned to n -gram matches	BLEU
geometric_mean	Geometric mean of n -gram precision scores	BLEU
W	Set of words in candidate caption	CIDEr
W_r	Set of words in reference captions	CIDEr
$freq(w, G)$	Term frequency of word w in G (entire corpus)	CIDEr
IDF(w)	Inverse Document Frequency of w	CIDEr
SPICE(S, R)	Graph-based semantic similarity between S (candidate) and R (reference)	SPICE

sentence and a set of reference sentences to measure their similarity at the sentence level. ROUGE-1, ROUGE-2, ROUGE-W, and ROUGE-SU4 are different types of ROUGE used for various tasks, and the metric score ranges from 0 to 1.

Calculating the longest common subsequence (LCS), the longest matching sequence of words between the original and predicted summaries, forms the basis of the ROUGE algorithm. Unlike matching words consecutively, LCS allows for matches that reflect the word order at the sentence level. Additionally, LCS automatically includes common n-grams in sequence, removing the need to calculate predetermined n -gram sequences. Mathematically, ROUGE can be defined as

$$F_{lcs} = \frac{(1 + \beta^2) \cdot R_{lcs} \cdot P_{lcs}}{R_{lcs} + \beta^2 \cdot P_{lcs}}. \quad (12)$$

The LCS-based precision P_{lcs} and the LCS-based recall R_{lcs} can be calculated using the upper part of (13) and (14) for the sentence level, or can be calculated using the lower part of the same equations for the summary level.

$$P_{lcs} = \left\{ \begin{array}{l} \frac{l_{LCS}(X, Y)}{\sum_{j=1}^{m_r} l_{LCS}(r_j, c)} \\ \frac{m_c}{m_r} \end{array} \right\} \quad (13)$$

$$R_{lcs} = \left\{ \begin{array}{l} \frac{l_{LCS}(X, Y)}{\sum_{j=1}^{m_r} l_{LCS}(r_j, c)} \\ \frac{m_r}{m_c} \end{array} \right\} \quad (14)$$

where LCS is the longest common subsequence, P_{lcs} : LCS-based precision, R_{lcs} : LCS-based recall, β : P_{lcs}/R_{lcs} , l_{LCS} : length of the longest common subsequence of X and Y , LCS U (r_j, c): LCS score of the union's longest common subsequence between a reference sentence and the candidate sentence.

6.3. The Metric for Explicit Ordering Translation Evaluation METEOR

The METEOR metric [126] is designed to evaluate machine translation and is considered more valuable than the Blue metric. Its correlation with human evaluations is stronger. Meteор calculates a score by comparing a candidate sentence with

a human-written reference sentence using generalized unigram matching. The score is computed based on the matched words' precision, recall, and alignment. When multiple reference sentences are involved, the candidate's final evaluation score is determined by choosing the best score among all independently computed ones. METEOR incorporates stemming, synonym matching, and standard exact word matching, making it more effective at the sentence or segment level [7]. The maximum score can be estimated by computing the F-measure through explicit unigram matching (i.e., word-for-word matching) between the candidate and reference translations. The METEOR metric is defined as

$$METEOR = F_{\text{mean}} \cdot (1 - pn) \quad (15)$$

The chunks comprise adjacent unigrams in the reference and hypothesis to calculate the penalty P_n . The longer the adjacent mappings are between the candidate and the reference, the fewer chunks there are. The penalty is obtained by

$$pn = 0.5 * \left(\frac{Ch}{U_m} \right)^3 \quad (16)$$

A harmonic mean of precision and recall is determined as the F-mean, with a higher value on recall as

$$F_{\text{mean}} = \frac{10.P.R}{R + 9.P} \quad (17)$$

and recall value R as

$$R = \frac{M(c)}{U(r)}, \quad (18)$$

and precision P as

$$P = \frac{M(c)}{U(c)} \quad (19)$$

where P_n is the penalty, Ch is the number of chunks, U_m is the number of unigrams that correspond between the candidate and the reference, $M(c)$ is the number of unigrams in the candidate sentence that are mapped, $U(r)$ is the total number of unigrams in the reference sentence, and $U(c)$ is the total number of unigrams in the candidate sentence.

6.4. Consensus-based Image Description Evaluation CIDEr

The Image Description Evaluation (IDE) tool uses the consensus-based Image Description Evaluation (CIDEr) metric to assess the similarity of a generated sentence to a set of human-authored ground truth sentences [127]. It employs a Term Frequency-Inverse Document Frequency (TF-IDF) weighting for each n-gram in the candidate phrase to encode their frequency in the reference sentences. This metric evaluates the grammar, relevance, and accuracy.

CIDEr was specifically designed to evaluate image captions and descriptions. Unlike other metrics that only work with five captions per image, it utilizes consensus through TF-IDF, making it unsuitable for analyzing the consensus between generated captions and human assessments [7]. Therefore, the average cosine similarity between the candidate and reference sentences is used to calculate the CIDEr score for n -grams of length n as

$$\text{CIDEr}_n(c, r_j) = \frac{1}{u_r} \sum_{j=1}^{u_r} \frac{g^n(c) \cdot g^n(r_j)}{\|g^n(c)\| \cdot \|g^n(r_j)\|} \quad (20)$$

The weighting TF-IDF $g_k(r_j)$ for each n -gram w_k of a reference sentence is defined as

$$g_k(r_j) = \frac{h_k(r_j)}{\sum_{w_l \in \Omega} h_l(r_j)} \log \left(\frac{|I|}{\sum_{I_p \in I} \min(1, \sum_j h_k(r_j))} \right) \quad (21)$$

Similarly, for $g_k(c)$, the candidate sentence is replaced by r_j with c . CIDEr is computed by combining the scores from n -grams of varying lengths as

$$\text{CIDEr}(c, r_j) = \sum_{n=1}^N w_n \cdot \text{CIDEr}_n(c, r_j) \quad (22)$$

where $g_n(c)$ is a vector formed by all n -grams of length n of the candidate sentence, $\|g_n(c)\|$ is the magnitude of the vector $g_n(c)$, $g_n(r_j)$ is a vector formed by all n -grams of length n of the set of reference sentences, $\|g_n(r_j)\|$ is the magnitude of the vectors $g_n(r_j)$, $g_k(r_j)$ is TF-IDF weighting for each n -gram w_k of the set of reference sentences, $g_k(c)$ is TF-IDF weighting for each n -gram w_k of the candidate sentence, $h_k(r_j)$ is the number of occurrences of an n -gram w_k in a reference sentence, $h_k(c)$ is the number of occurrences of an n -gram w_k in the candidate sentence, Ω is the vocabulary of all n -grams, and I is the set of all images in the dataset.

6.5. The Semantic Propositional Image Caption Evaluation SPICE

SPICE (Semantic Propositional Image Caption Evaluation) was developed to evaluate image captioning using semantic scene

graphs [128]. It is considered to be more accurate than human judgments. The process involves extracting information about various items, properties, and their relationships from image descriptions [2]. The captions are converted into scene graphs via semantic parsing. The similarity score between the generated and ground-truth caption scene graphs is calculated using precision and recall F1 scores.

Precision is determined by the matching tuples between the logical tuples for generated and reference captions divided by the total number of logical tuples in the generated caption set. For recall, the matching tuples are divided by the total number of logical tuples in the reference caption set. The F1 score (SPICE) is calculated using precision and recall [27]. Scene graphs ($G(c)$ and $G(S_r)$) are created from candidate and reference captions, respectively), and the F score is calculated using the conjunction of logical tuples representing semantic propositions in the scene graph. SPICE can be calculated using the scene graphs of all reference sentences as

$$\text{SPICE}(c, S_r) = F_1(c, S_r) = \frac{2 \cdot P(c, S_r) \cdot R(c, S_r)}{P(c, S_r) + R(c, S_r)} \quad (23)$$

Precision and recall are calculated as in (24) and (25), respectively.

$$P = \frac{|T(G(c)) \otimes T(G(S_r))|}{|T(G(c))|} \quad (24)$$

$$R = \frac{|T(G(c)) \otimes T(G(S_r))|}{|T(G(S_r))|} \quad (25)$$

where $G(c)$ is the scene graph of the candidate sentence, $G(r_j)$ is the scene graph of each reference sentence, $G(S_r)$ is the scene graph of all reference sentences, $O(c)$ is set of objects in the candidate sentence, $E(c)$ is set of attributes in the candidate sentence, $K(c)$ is set of relations in the candidate sentence, T is the function that allows us to return logical tuples.

7. Limitation and Challenges

The development of image captioning models faces several challenges, including exploding gradients and the generation of incorrect sentences. Most modern algorithms rely on Recurrent Neural Networks and Long-Short-Term Memory Networks, which can suffer from vanished gradients and require significant resources, making them less hardware-friendly [84]. Although Generative Adversarial Networks [97] offer a promising alternative, they come with their own set of issues, such as the difficulty of training due to the discrete nature of GAN [129, 130]. Another approach involves using semantic feature vectors or focusing on object-region relationships [131] or focusing on object-region relationships [6] [53] [5].

Current methods for evaluating caption quality use logarithmic likelihood scores and automated metrics such as BLEU, METEOR, ROUGE, CIDEr, and SPICE. However, these metrics often do not correlate well with human evaluations. Despite SPICE and CIDEr being closer to human judgment, they still

challenge optimization [2]. These automated measures mainly focus on lexical or semantic data and do not fully capture the complex relationships between words and objects [46]. To enhance captioning models, it is essential to improve automatic evaluation methods to more accurately reflect human judgments and address gaps in understanding object and word relationships [103].

As image captioning technology advances, several future challenges must be addressed to enhance its effectiveness. One major challenge is improving the adaptability of captioning systems to diverse languages and cultural contexts. Current models often struggle with non-English languages and culturally nuanced images, leading to less accurate or contextually relevant captions. Another challenge is the need for more comprehensive and inclusive datasets. Many existing datasets are limited in scope or do not represent a wide range of cultural and contextual variations, affecting the generalizability of captioning systems. Additionally, there is a growing need to develop models that can understand and generate captions for complex, abstract, or ambiguous images, where traditional methods may fall short. Ensuring these systems can operate effectively in real-world scenarios, including understanding user-specific contexts and preferences, is also crucial.

8. Conclusion and Research Opportunities

Image captioning has emerged as a crucial task at the intersection of computer vision and natural language processing, enabling machines to understand and describe visual content. This survey has comprehensively reviewed attention-based transformer models, covering their architectures, evaluation metrics, datasets, and multilingual applications. We highlighted the transition from traditional template-based approaches to deep learning-driven transformer models, emphasizing the role of attention mechanisms in improving caption quality. Despite significant advancements, key challenges include handling complex scene compositions, improving caption fluency in low-resource languages, and ensuring factual accuracy in generated descriptions. Future research opportunities could focus on:

- a) Multimodal learning and cross-domain image captioning: Integrating vision, language, and other sensory inputs for richer, more context-aware captions.
- b) Multilingual and cross-lingual captioning: Expanding datasets and improving transfer learning techniques for non-English languages.
- c) Real-time and interactive captioning: Optimizing models for assistive AI, augmented reality, and robotics applications.
- d) Applications in novel domains: Extending image captioning to forensic analysis, cultural heritage, and personalized AI assistants.

As AI evolves, attention-based transformer models will remain superior in bridging the gap between vision and language. Addressing the challenges and opportunities outlined in this survey will be essential for realizing the full potential of image captioning in diverse real-world applications.

Abbreviations

The abbreviations used in this manuscript are given in Table 5.

Table 5: List of abbreviations

Abbreviation	Full Form
AI	Artificial Intelligence
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Unit
NLP	Natural Language Processing
CV	Computer Vision
GAN	Generative Adversarial Network
BLEU	Bilingual Evaluation Understudy (Metric)
METEOR	Metric for Evaluation of Translation with Explicit ORdering
CIDEr	Consensus-based Image Description Evaluation
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
SPICE	Semantic Propositional Image Caption Evaluation
VLP	Vision-Language Pre-training
MS COCO	Microsoft Common Objects in Context (Dataset)
SLR	Systematic Literature Review

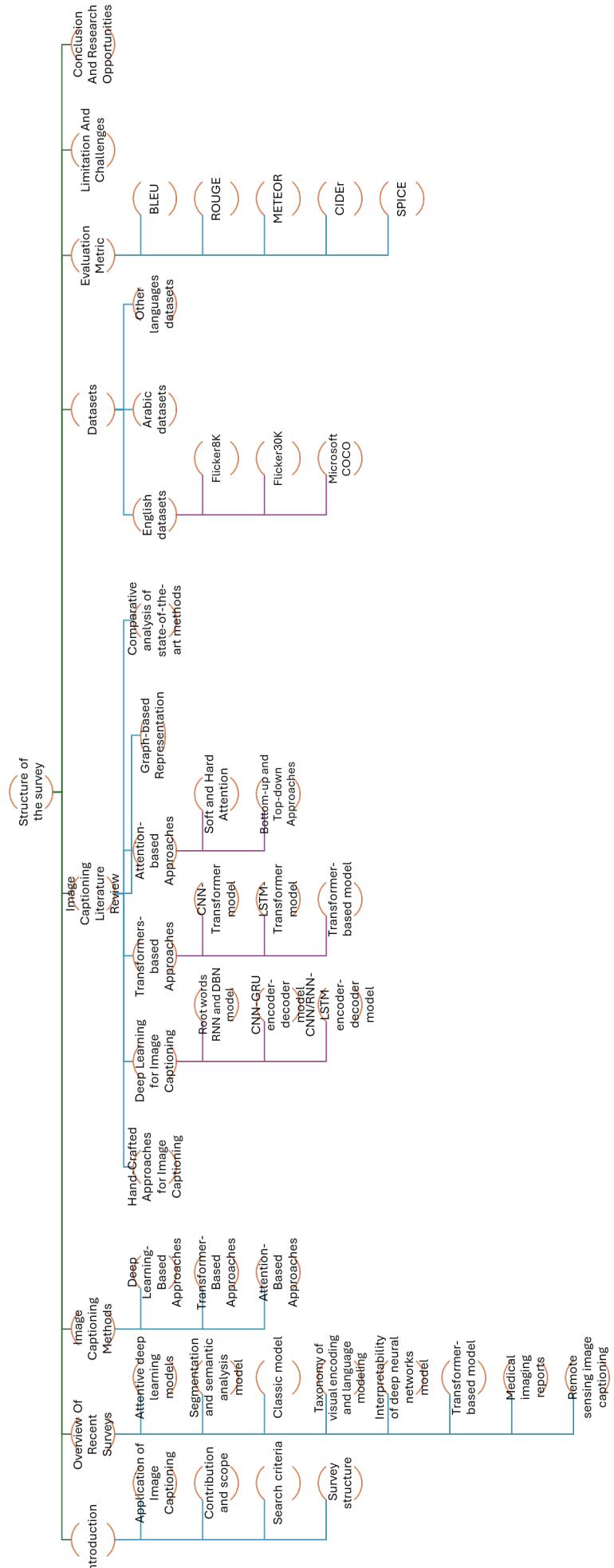


Table 6: Summary of multilingual image captioning models

Reference	Image model	Language model	Dataset	Dataset's language	Improvement
Zhang 2021 [45]	LSTM	LSTM	Flickr30k	English	Provided fine-grained information among objects
AI 2018 [56]	VGG	RNN-LSTM	MSCOCO, Flickr8k	Arabic	The model can achieve excellent results with larger corpus
Chen 2019 [93]	CNN	RNN	Daily Mail	English	Considering both the news; image and text
Biswas 2020 [6]	CNN	LSTM	MSCOCO	English	Provided further improvements in image captioning
Mualla 2018 [57]	CNN	LSTM	Flickr8k	Arabic	The English-based model had the best performance
Eljundi 2020 [58]	VGG16	LSTM layer	Flickr8K	Arabic	Showed the superiority of the end-to-end model
Do 2020 [69]	CNN	GRU	Flickr30K MSCOCO	English	PoS gave the best performance
Saleh 2019 [95]	CNN	LSTM	MSCOCO	Arabic	Convert stories to images support the meaning
Cheikh 2020 [4]	CNN	LSTM	ArabicFlickr1K	Arabic	Applied human annotators
Mulyanto 2019 [81]	CNN	LSTM	Flickr	Indonesian	The test set provided promising results
Tien 2020 [3]	CNN	bi-RNN	MSCOCO	Vietnamese	A solution to the problem of unknown words
Pa 2020 [84]	CNN	LSTM	Flickr8k	Myanmar	Automatic translation reduced the manual captioning time
Yu 2019 [5]	R-CNN	LSTM layer	MSCOCO	English	Three types of relations
He 2020 [47]	Transformer	Transformer	MSCOCO	English	Query region: parent, neighbor, or child
Pedersoli 2017 [53]	CNN	GRU	MSCOCO	English	Modeled the direct dependencies between words and image
Fei 2022 [77]	Transformer	Transformer	MSCOCO	English	Image region features to direct attention alignment
WEI 2022 [104]	RNN	Transformer	MSCOCO	English	Capture the dependencies within the image areas and regions
Wang 2022 [101]	LSTM	Transformer	MSCOCO Flickr30k	English	Utilizes geometry relations
Yan 2022 [115]	FR-CNN	Transformer	MSCOCO	English	Reduces the misinformation
Lu 2023 [49]	Transformer	Transformer	MSCOCO	English	Common self-attentive mechanism
Dubey 2023 [105]	FRCNN	Transformer	MSCOCO	English	Relate objects based on localized ratios
Wang 2021 [116]	FRCNN	Transformer	MSCOCO	English	Dynamic attention aggregation
Wang 2020 [50]	R-CNN	Transformer	MSCOCO	English	Capture critical objects and relationships
Guo 2020 [98]	FRCNN	Transformer	MSCOCO	English	Provided a unique normalization method
Jiang 2021 [106]	FR-CNN	LSTM	MSCOCO	English	Simplify the model and increase efficiency
Liu 2022 [117]	Transformer	Transformer	MSCOCO	English	Visual attention and keyword attention
Liu 2022 [102]	ResNet50	Transformer	RSICD	English	Effectively construct sentences
Kandala 2022 [107]	Transformer	Transformer	UC-Merced	English	Self-attention improve the performance
Nguyen 2022 [100]	Transformer	Transformer	MSCOCO	English	Transformer get beyond the drawback of CNN
Tan 2022 [108]	Transformer	Transformer	MSCOCO	English	Fewer parameters without loss of performance
Kumar 2022 [99]	Inception-V3	Transformer	Flickr MSCOCO,	English	intra- and inter-model interactions
Wang 2022 [109]	Transformer	Transformer	MSCOCO	English	high-order intra- and inter-feature interactions
Ma 2023 [67]	Transformer	Transformer	Flickr MSCOCO	English	Local visual modeling with grid features

References

- [1] S. Bai, S. An, A survey on automatic image caption generation, *Neurocomputing* 311 (2018) 291–304.
- [2] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, H. Laga, A comprehensive survey of deep learning for image captioning, *ACM Computing Surveys (CSUR)* 51 (6) (2019) 1–36.
- [3] H. N. Tien, T.-H. Do, V.-A. Nguyen, Image captioning in vietnamese language based on deep learning network, in: *International Conference on Computational Collective Intelligence*, Springer, 2020, pp. 789–800.
- [4] M. Cheikh, M. Zrigui, Active learning based framework for image captioning corpus creation, in: *International Conference on Learning and Intelligent Optimization*, Springer, 2020, pp. 128–142.
- [5] J. Yu, J. Li, Z. Yu, Q. Huang, Multimodal transformer with multi-view visual representation for image captioning, *IEEE transactions on circuits and systems for video technology* 30 (12) (2019) 4467–4480.
- [6] R. Biswas, M. Barz, D. Sonntag, Towards explanatory interactive image captioning using top-down and bottom-up features, beam search and re-ranking, *KI-Künstliche Intelligenz* 34 (4) (2020) 571–584.
- [7] T. Ghandi, H. Pourreza, H. Mahyar, Deep learning approaches on image captioning: A review, *ACM Comput. Surv.* 56 (3) (oct 2023).
- [8] O. S. Al-Kadi, Combined statistical and model based texture features for improved image classification, in: *4th IET International Conference on Advances in Medical, Signal and Information Processing-MEDSIP 2008*, IET, 2008, pp. 1–4.
- [9] O. S. Al-Kadi, Supervised texture segmentation: a comparative study, in: *2011 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*, IEEE, 2011, pp. 1–5.
- [10] H. Ayesha, S. Iqbal, M. Tariq, M. Abrar, M. Sanaullah, I. Abbas, A. Rehman, M. F. K. Niazi, S. Hussain, Automatic medical image interpretation: State of the art and future directions, *Pattern Recognition* (2021) 107856.
- [11] P. Chendake, P. Korpal, S. Bhor, R. Bansal, S. Patil, D. Deshpande, Learning system for kids, *International Journal of Recent Advances in Multidisciplinary Topics* 2 (6) (2021) 71–75.
- [12] A. Ogura, N. Hayashi, T. Negishi, H. Watanabe, Effectiveness of an e-learning platform for image interpretation education of medical staff and students, *Journal of digital imaging* 31 (5) (2018) 622–627.
- [13] A. K. Muhammed Kunju, S. Baskar, S. Zafar, B. AR, A transformer based real-time photo captioning framework for visually impaired people with visual attention, *Multimedia Tools and Applications* (2024) 1–20.
- [14] D. H. Fudholi, Y. Windiatmoko, N. Afrianto, P. E. Susanto, M. Suyuti, A. F. Hidayatullah, R. Rahmadi, Image captioning with attention for smart local tourism using efficientnet, in: *IOP Conference Series: Materials Science and Engineering*, Vol. 1077, IOP Publishing, 2021, p. 012038.
- [15] M. Nivedita, P. Chandrashekhar, S. Mahapatra, Y. A. V. Phamila, S. K. Selvaperumal, Image captioning for video surveillance system using neural networks, *International Journal of Image and Graphics* (2021) 2150044.
- [16] G. Hoxha, F. Melgani, B. Demir, Toward remote sensing image retrieval under a deep image captioning perspective, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 13 (2020) 4462–4475.
- [17] Z. Wang, Z. Huang, Y. Luo, Paic: Parallelised attentive image captioning, in: *Australasian Database Conference*, Springer, 2020, pp. 16–28.
- [18] K. Shuster, S. Humeau, H. Hu, A. Bordes, J. Weston, Engaging image captioning via personality, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12516–12526.
- [19] H. Fujiyoshi, T. Hirakawa, T. Yamashita, Deep learning-based image recognition for autonomous driving, *IATSS research* 43 (4) (2019) 244–252.
- [20] A. P. Shah, J.-B. Lamare, T. Nguyen-Anh, A. Hauptmann, Cadp: A novel dataset for cctv traffic camera based accident analysis, in: *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, IEEE, 2018, pp. 1–9.
- [21] D. Guinness, E. Cutrell, M. R. Morris, Caption crawler: Enabling reusable alternative text descriptions using reverse image search, in: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–11.
- [22] Q. Huang, L. Yang, H. Huang, T. Wu, D. Lin, Caption-supervised face recognition: Training a state-of-the-art face model without manual annotation, in: *European Conference on Computer Vision*, Springer, 2020, pp. 139–155.
- [23] P. P. Khaing, et al., Attention-based deep learning model for image captioning: a comparative study, *International Journal of Image, Graphics and Signal Processing* 11 (6) (2019) 1.
- [24] F. Chen, X. Li, J. Tang, S. Li, T. Wang, A survey on recent advances in image captioning, in: *Journal of Physics: Conference Series*, Vol. 1914, IOP Publishing, 2021, p. 012053.
- [25] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, R. Cucchiara, From show to tell: a survey on deep learning-based image captioning, *IEEE transactions on pattern analysis and machine intelligence* 45 (1) (2022) 539–559.
- [26] G. Luo, L. Cheng, C. Jing, C. Zhao, G. Song, A thorough review of models, evaluation metrics, and datasets on image captioning, *IET Image Processing* 16 (2) (2022) 311–332.
- [27] Z. Zohourianshahzadi, J. K. Kalita, Neural attention for image captioning: review of outstanding methods, *Artificial Intelligence Review* (2021) 1–30.
- [28] H. Senior, G. Slabaugh, S. Yuan, L. Rossi, Graph neural networks in vision-language image understanding: a survey, *The Visual Computer* (2024) 1–26.
- [29] T. Pang, P. Li, L. Zhao, A survey on automatic generation of medical imaging reports based on deep learning, *BioMedical Engineering On-Line* 22 (1) (2023) 48.
- [30] C. Wohlin, Guidelines for snowballing in systematic literature studies and a replication in software engineering, in: *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering, EASE ’14*, Association for Computing Machinery, New York, NY, USA, 2014, pp. 1–10.
- [31] Z. Zohourianshahzadi, J. K. Kalita, Neural attention for image captioning: review of outstanding methods, *Artificial Intelligence Review* 55 (5) (2022) 3833–3862.
- [32] H. Sharma, D. Padha, A comprehensive survey on image captioning: from handcrafted to deep learning-based techniques, a taxonomy and open research issues, *Artificial Intelligence Review* 56 (11) (2023) 13619–13661.
- [33] H. Sharma, D. Padha, Domain-specific image captioning: a comprehensive review, *International Journal of Multimedia Information Retrieval* 13 (2) (2024) 20.
- [34] H. Sharma, D. Padha, A. Selwal, A survey on attention-based image captioning: Taxonomy, challenges, and future perspectives, in: *International Conference on Machine Intelligence and Signal Processing*, Springer, 2022, pp. 681–694.
- [35] A. Oluwasammi, M. U. Aftab, Z. Qin, S. T. Ngo, T. V. Doan, S. B. Nguyen, S. H. Nguyen, G. H. Nguyen, Features to text: a comprehensive survey of deep learning on semantic segmentation and image captioning, *Complexity* 2021 (2021).
- [36] R. Staniūtė, D. Šešok, A systematic literature review on image captioning, *Applied Sciences* 9 (10) (2019) 2024.
- [37] H. Sharma, D. Padha, From templates to transformers: a survey of multimodal image captioning decoders, in: *2023 International Conference on Computer, Electronics & Electrical Engineering & their Applications (IC2E3)*, IEEE, 2023, pp. 1–6.
- [38] X. Liu, Q. Xu, N. Wang, A survey on deep neural network-based image captioning, *The Visual Computer* 35 (3) (2019) 445–470.
- [39] S.-H. Choi, S. Y. Jo, S. H. Jung, Component based comparative analysis of each module in image captioning, *ICT Express* 7 (1) (2021) 121–125.
- [40] R. Thirunavukarasu, E. Kotei, A comprehensive review on transformer network for natural and medical image analysis, *Computer Science Review* 53 (2024) 100648. doi:<https://doi.org/10.1016/j.cosrev.2024.100648>. URL <https://www.sciencedirect.com/science/article/pii/S1574013724000137>
- [41] E. Kotei, R. Thirunavukarasu, Medical image analysis with vision transformers for downstream tasks and clinical report generation, in: *Intelligent Systems and Sustainable Computational Models*, Auerbach Publications, pp. 288–307.
- [42] B. Zhao, A systematic survey of remote sensing image captioning, *IEEE Access* 9 (2021) 154086–154111.
- [43] X. Lu, B. Wang, X. Zheng, X. Li, Exploring models and data for remote

- sensing image caption generation, *IEEE Transactions on Geoscience and Remote Sensing* 56 (4) (2017) 2183–2195.
- [44] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, H. Laga, M. Bennamoun, Text to image synthesis for improved image captioning, *IEEE Access* 9 (2021) 64918–64928.
- [45] W. Zhang, H. Shi, S. Tang, J. Xiao, Q. Yu, Y. Zhuang, Consensus graph representation learning for better grounded image captioning, in: Proc 35 AAAI Conf on Artificial Intelligence, 2021, pp. 3394–3402.
- [46] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in neural information processing systems, 2017, pp. 5998–6008.
- [47] S. He, W. Liao, H. R. Tavakoli, M. Yang, B. Rosenhahn, N. Pugeault, Image captioning through image transformer, in: Proceedings of the Asian Conference on Computer Vision, 2020, pp. 153–169.
- [48] T. Wolf, J. Chaumond, L. Debut, V. Sanh, C. Delangue, A. Moi, P. Cistac, M. Funtowicz, J. Davison, S. Shleifer, et al., Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 38–45.
- [49] T. Lu, J. Wang, F. Min, Full-memory transformer for image captioning, *Symmetry* 15 (1) (2023) 190.
- [50] D. Wang, H. Hu, D. Chen, Transformer with sparse self-attention mechanism for image captioning, *Electronics Letters* 56 (15) (2020) 764–766.
- [51] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: International conference on machine learning, PMLR, 2015, pp. 2048–2057.
- [52] H. Wang, Y. Zhang, X. Yu, An overview of image caption generation methods, *Computational intelligence and neuroscience* 2020 (2020).
- [53] M. Pedersoli, T. Lucas, C. Schmid, J. Verbeek, Areas of attention for image captioning, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 1242–1250.
- [54] V. Jindal, A deep learning approach for arabic caption generation using roots-words, in: Thirty-First AAAI Conference on Artificial Intelligence, 2017, pp. 4941–4942.
- [55] V. Jindal, Generating image captions in arabic using root-word based recurrent neural networks and deep neural networks, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, 2018, p. 144–151.
- [56] H. A. Al-Muzaini, T. N. Al-Yahya, H. Benhidour, Automatic arabic image captioning using rnn-lst m-based language model and cnn, *International Journal of Advanced Computer Science and Applications* 9 (6) (2018).
- [57] R. Mualla, J. Alkheir, Development of an arabic image description system, *International Journal of Computer Science Trends and Technology (IJCST)*–6 (3) (2018) 205–213.
- [58] O. ElJundi, M. Dhaybi, K. Mokadam, H. M. Hajj, D. C. Asmar, Resources and end-to-end neural network models for arabic image captioning., in: VISIGRAPP (5: VISAPP), 2020, pp. 233–241.
- [59] H. Hejazi, K. Shaalan, Deep learning for arabic image captioning: A comparative study of main factors and preprocessing recommendations, *International Journal of Advanced Computer Science and Applications* 12 (11) (2021).
- [60] S. M. Sabri, Arabic image captioning using deep learning with attention, Ph.D. thesis, University of Georgia (2021).
- [61] J. Emami, P. Nugues, A. Elnagar, I. Afyouni, Arabic image captioning using pre-training of deep bidirectional transformers, in: Proceedings of the 15th International Conference on Natural Language Generation, 2022, pp. 40–51.
- [62] M. T. Lasheen, N. H. Barakat, Arabic image captioning: the effect of text pre-processing on the attention weights and the bleu-n scores, *Int J Adv Comput Sci Appl* 13 (7) (2022) 11.
- [63] A. Alsayed, T. M. Qadah, M. Arif, A performance analysis of transformer-based deep learning models for arabic image captioning, *Journal of King Saud University-Computer and Information Sciences* 35 (9) (2023) 101750.
- [64] S. Elbedwehy, T. Medhat, Improved arabic image captioning model using feature concatenation with pre-trained word embedding, *Neural Computing and Applications* 35 (2023) 1–17. doi:10.1007/s00521-023-08744-1.
- [65] A. Karpathy, 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.
- [66] J. Bineeshia, Image caption generation using cnn-lstm based approach, in: Proceedings of the First International Conference on Combinatorial and Optimization, ICCAP 2021, December 7–8 2021, Chennai, India, 2021, pp. 1–9.
- [67] Y. Ma, J. Ji, X. Sun, Y. Zhou, R. Ji, Towards local visual modeling for image captioning, *Pattern Recognition* 138 (2023) 109420.
- [68] T. Jiang, Z. Zhang, Y. Yang, Modeling coverage with semantic embedding for image caption generation, *The Visual Computer* 35 (11) (2019) 1655–1665.
- [69] T. do Carmo Nogueira, C. D. N. Vinhal, G. da Cruz Júnior, M. R. D. Ullmann, Reference-based model using multimodal gated recurrent units for image captioning, *Multimedia Tools and Applications* 79 (41) (2020) 30615–30635.
- [70] J. Lu, C. Xiong, D. Parikh, 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.
- [71] M. Kalimuthu, A. Mogadala, M. Mosbach, D. Klakow, Fusion models for improved image captioning, in: Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part VI, Springer, 2021, pp. 381–395.
- [72] A. Abdussalam, Z. Ye, A. Hawbani, M. Al-Qatf, R. Khan, Numcap: a number-controlled multi-caption image captioning network, *ACM Transactions on Multimedia Computing, Communications and Applications* 19 (4) (2023) 1–24.
- [73] A. Shrimai, T. Chakraborty, Attention beam: An image captioning approach (student abstract), in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 15887–15888.
- [74] W. Zhao, X. Wu, J. Luo, Cross-domain image captioning via cross-modal retrieval and model adaptation, *IEEE Transactions on Image Processing* 30 (2020) 1180–1192.
- [75] C. Wang, Y. Shen, L. Ji, Geometry attention transformer with position-aware lstms for image captioning, *Expert Systems with Applications* 201 (2022) 117174.
- [76] H. Zhu, R. Wang, X. Zhang, Image captioning with dense fusion connection and improved stacked attention module, *Neural Process. Lett.* 53 (2) (2021) 1101–1118.
- [77] Z. Fei, Attention-aligned transformer for image captioning, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, pp. 607–615.
- [78] Y. Wang, J. Xu, Y. Sun, End-to-end transformer based model for image captioning, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, pp. 2585–2594.
- [79] X. Yang, Y. Liu, X. Wang, Reformer: The relational transformer for image captioning, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 5398–5406.
- [80] R. Mulyawan, A. Sunyoto, A. H. Muhammad, Automatic indonesian image captioning using cnn and transformer-based model approach, in: 2022 5th International Conference on Information and Communications Technology (ICOIACT), 2022, pp. 355–360.
- [81] E. Mulyanto, E. I. Setiawan, E. M. Yuniarso, M. H. Purnomo, Automatic indonesian image caption generation using cnn-lstm model and feeh-id dataset, in: 2019 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA), IEEE, 2019, pp. 1–5.
- [82] J. J. Wijadi, B. Ghiffar, N. N. Qomariyah, Indonesian language image captioning using encoder-decoder with attention approach, in: 2024 2nd International Conference on Software Engineering and Information Technology (ICoSEIT), 2024, pp. 342–347.
- [83] A. A. Nugraha, A. Arifianto, Suyanto, Generating image description on indonesian language using convolutional neural network and gated recurrent unit, 2019 7th International Conference on Information and Communication Technology (ICoICT) (2019) 1–6.
- [84] W. P. Pa, T. L. Nwe, et al., Automatic myanmar image captioning using cnn and lstm-based language model, in: Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL), 2020, pp. 139–143.
- [85] S. P. P. Aung, W. P. Pa, T. L. Nwe, Improving myanmar image cap-

- tion generation using nasnetlarge and bi-directional lstm, in: 2023 IEEE Conference on Computer Applications (ICCA), 2023, pp. 1–6.
- [86] F. Muhammad Shah, M. Humaira, M. A. R. K. Jim, A. Saha Ami, S. Paul, Bornon: Bengali image captioning with transformer-based deep learning approach, *SN Computer Science* 3 (2022) 1–16.
- [87] M. F. Khan, S. Sadiq-Ur-Rahman, M. S. Islam, Improved bengali image captioning via deep convolutional neural network based encoder-decoder model, in: Proceedings of International Joint Conference on Advances in Computational Intelligence, Springer, 2021, pp. 217–229.
- [88] M. Humaira, P. Shimul, M. A. R. K. Jim, A. S. Ami, F. M. Shah, A hybridized deep learning method for bengali image captioning, *International Journal of Advanced Computer Science and Applications* 12 (2) (2021).
- [89] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, T. L. Berg, Babytalk: Understanding and generating simple image descriptions, *IEEE transactions on pattern analysis and machine intelligence* 35 (12) (2013) 2891–2903.
- [90] K. Kpalma, J. Ronsin, An overview of advances of pattern recognition systems in computer vision, *Vision Systems* (2007) 26.
- [91] M. Sezgin, B. I. Sankur, Survey over image thresholding techniques and quantitative performance evaluation, *Journal of Electronic imaging* 13 (1) (2004) 146–168.
- [92] W. Almanaseer, M. Alshraideh, O. Al-Kadi, A deep belief network classification approach for automatic diacritization of arabic text, *Applied Sciences* 11 (11) (2021) 5228.
- [93] J. Chen, H. Zhuge, A news image captioning approach based on multimodal pointer-generator network, *Concurrency and Computation: Practice and Experience* (2019) e5721.
- [94] J. Zakraoui, S. Elloumi, J. M. Alja'am, S. B. Yahia, Improving arabic text to image mapping using a robust machine learning technique, *IEEE Access* 7 (2019) 18772–18782.
- [95] M. Saleh, J. M. Alja'am, Towards adaptive multimedia system for assisting children with arabic learning difficulties, in: 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), IEEE, 2019, pp. 794–799.
- [96] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, L. Farhan, Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions, *Journal of big Data* 8 (1) (2021) 1–74.
- [97] W. Zhang, W. Nie, X. Li, Y. Yu, Image caption generation with adaptive transformer, in: 2019 34rd Youth Academic Annual Conference of Chinese Association of Automation (YAC), IEEE, 2019, pp. 521–526.
- [98] L. Guo, J. Liu, X. Zhu, P. Yao, S. Lu, 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. 10327–10336.
- [99] D. Kumar, V. Srivastava, D. E. Popescu, J. D. Hemanth, Dual-modal transformer with enhanced inter-and intra-modality interactions for image captioning, *Applied Sciences* 12 (13) (2022) 6733.
- [100] V.-Q. Nguyen, M. Suganuma, T. Okatani, Grit: Faster and better image captioning transformer using dual visual features, in: Computer Vision-ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI, Springer, 2022, pp. 167–184.
- [101] C. Wang, Y. Shen, L. Ji, Geometry attention transformer with position-aware lstms for image captioning, *Expert Systems with Applications* 201 (2022) 117174.
- [102] C. Liu, R. Zhao, Z. Shi, Remote-sensing image captioning based on multilayer aggregated transformer, *IEEE Geoscience and Remote Sensing Letters* 19 (2022) 1–5.
- [103] G. Li, L. Zhu, P. Liu, Y. Yang, Entangled transformer for image captioning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 8928–8937.
- [104] Y. Wei, C. Wu, G. Li, H. Shi, Sequential transformer via an outside-in attention for image captioning, *Engineering Applications of Artificial Intelligence* 108 (2022) 104574.
- [105] S. Dubey, F. Olimov, M. A. Rafique, J. Kim, M. Jeon, Label-attention transformer with geometrically coherent objects for image captioning, *Information Sciences* 623 (2023) 812–831.
- [106] W. Jiang, X. Li, H. Hu, Q. Lu, B. Liu, Multi-gate attention network for image captioning, *IEEE Access* 9 (2021) 69700–69709.
- [107] H. Kandala, S. Saha, B. Banerjee, X. X. Zhu, Exploring transformer and multilabel classification for remote sensing image captioning, *IEEE Geoscience and Remote Sensing Letters* 19 (2022) 1–5.
- [108] J. H. Tan, Y. H. Tan, C. S. Chan, J. H. Chuah, Acort: A compact object relation transformer for parameter efficient image captioning, *Neurocomputing* 482 (2022) 60–72.
- [109] X. Wang, X. Fang, Y. Yang, Dm-catn: Deep modular co-attention transformer networks for image captioning, in: International Conference on Artificial Intelligence and Intelligent Information Processing (AIIIP 2022), Vol. 12456, SPIE, 2022, pp. 600–606.
- [110] M. Cornia, M. Stefanini, L. Baraldi, R. Cucchiara, Meshed-memory transformer for image captioning, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10578–10587.
- [111] M. Cornia, L. Baraldi, R. Cucchiara, Explaining transformer-based image captioning models: An empirical analysis, *AI Communications* 35 (2) (2022) 111–129.
- [112] Y. Pan, T. Yao, Y. Li, T. Mei, X-linear attention networks for image captioning, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10971–10980.
- [113] L. Huang, W. Wang, J. Chen, X.-Y. Wei, Attention on attention for image captioning, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 4634–4643.
- [114] L. Ke, W. Pei, R. Li, X. Shen, Y.-W. Tai, Reflective decoding network for image captioning, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 8888–8897.
- [115] C. Yan, Y. Hao, L. Li, J. Yin, A. Liu, Z. Mao, Z. Chen, X. Gao, Task-adaptive attention for image captioning, *IEEE Transactions on Circuits and Systems for Video Technology* 32 (1) (2022) 43–51.
- [116] Y. Wang, X. Sun, X. Li, W. Zhang, X. Gao, Reasoning like humans: on dynamic attention prior in image captioning, *Knowledge-Based Systems* 228 (2021) 107313.
- [117] J. Liu, K. Cheng, H. Jin, Z. Wu, An image captioning algorithm based on combination attention mechanism, *Electronics* 11 (9) (2022) 1397.
- [118] H. Li, P. Wang, C. Shen, G. Zhang, Show, attend and read: A simple and strong baseline for irregular text recognition, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 33, 2019, pp. 8610–8617.
- [119] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, 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.
- [120] T. Yao, Y. Pan, Y. Li, T. Mei, Exploring visual relationship for image captioning, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 684–699.
- [121] M. Hodosh, P. Young, J. Hockenmaier, Framing image description as a ranking task: Data, models and evaluation metrics, *Journal of Artificial Intelligence Research* 47 (2013) 853–899.
- [122] P. Young, A. Lai, M. Hodosh, J. Hockenmaier, From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions, *Transactions of the Association for Computational Linguistics* 2 (2014) 67–78.
- [123] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: European conference on computer vision, Springer, 2014, pp. 740–755.
- [124] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.
- [125] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text summarization branches out, 2004, pp. 74–81.
- [126] S. Banerjee, 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.
- [127] R. Vedantam, C. Lawrence Zitnick, D. Parikh, Cider: Consensus-based image description evaluation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 4566–4575.
- [128] P. Anderson, B. Fernando, M. Johnson, S. Gould, Spice: Semantic propositional image caption evaluation, in: European conference on computer vision, Springer, 2016, pp. 382–398.

- [129] A. Abu-Srhan, M. A. Abushariah, O. S. Al-Kadi, The effect of loss function on conditional generative adversarial networks, *Journal of King Saud University-Computer and Information Sciences* 34 (9) (2022) 6977–6988.
- [130] A. Abu-Srhan, I. Almallahi, M. A. Abushariah, W. Mahafza, O. S. Al-Kadi, Paired-unpaired unsupervised attention guided gan with transfer learning for bidirectional brain mr-ct synthesis, *Computers in Biology and Medicine* 136 (2021) 104763.
- [131] L. Gong, J. M. Crego, J. Senellart, Enhanced transformer model for data-to-text generation, in: *Proceedings of the 3rd Workshop on Neural Generation and Translation*, 2019, pp. 148–156.