

A Comprehensive Analysis of Image Caption Generation Using Multi-Deep-Neural Network Models

Debkumar Chowdhury^{1*}, Debosmita Chaudhuri², Jayanta Aich³, Susovon Chowdhury⁴ and Kartik Sau⁵

^{1*}Department of Computer Science and Engineering, Guru Nanak Institute of Technology, Nilgunj Road, Sodepur, 700114, West Bengal, India.

²Department of Computer Science and Engineering, Brainware University, 398, Ramkrishnapur Road, Barasat, Kolkata - 700125, West Bengal, India.

³Department of Computational Sciences, Brainware University, 398, Ramkrishnapur Road, Barasat, Kolkata - 700125, West Bengal, India.

⁴Director of Asphodel Education Point, Asphodel Education Point, Narayanpur, Balurghat, Dinajpur-733101, West Bengal, India.

⁵Department of Basic Science and Humanities, University of Engineering and Management, Kolkata, Plot No. III, B/5, New Town Rd, Action Area III, Newtown, Kolkata - 700160, West Bengal, India.

Abstract

This article delves into the realm of image caption generation, encompassing a thorough examination of its mechanisms, and constraints, a comprehensive portrayal of diverse datasets, an extensive elucidation of various performance metrics, and a detailed comparative analysis. To accomplish this undertaking, we scrutinized an array of image caption generation methodologies that have emerged over the past

decade. We also categorized these methodologies deeply, distinguishing between primitive approaches, deep neural network-based techniques, and attention-based methods. These categorizations were further detailed using algorithms, procedures, equations, and block diagrams. This paper presents both basic and advanced deep-learning algorithms for generating image captions. The datasets known as MS-COCO and Flickr are utilised here to evaluate the performance methods. The article presents an all-encompassing analysis of the existing methods through comparative tables and graphs, employing various quantitative parameters, including BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE, CIDEr, SPICE, P-Rate, R-Rate, F1 Score, and A-Score. By closely analyzing the quantitative parameters BLEU-1, BLEU-2, BLEU-3, BLEU-4, ROUGE, CIDEr, and SPICE on the MSCOCO and Flickr datasets, clearly the MADASAP achieves the highest scores, outperforming other models. When considering the METEOR parameter on the same datasets, the NICVATP2L and MADASAP models stand out, each generating the highest METEOR scores, respectively. Following this comparative analysis, it is clear that the GT+ SCD Method stands out in terms of P-Rate, R-Rate, and A-Rate. Meanwhile, the UCMA model achieves the highest F1 Score compared to other methods. This analysis provides researchers with valuable insights for developing robust techniques that address the limitations of existing approaches and incorporate the desirable attributes of current methods.

Keywords: Primitive Caption Generation Methodology, Generic Search Language Based Methodology, Deep Neural Network Based Methodology, Attention Based Methodology, Transformer Based Methodology, Graph-Based Methodology, CNN Based Methodology, Unsupervised Based Methodology, Reinforcement Learning Based Methodology, Encoder-Decoder Based Methodology

1 Introduction

Image caption generation involves identifying the context of the image and annotating appropriate captions using computer-based methods. This process labels an image with relevant keywords. Various datasets are often employed during model training to aid in accurate context labelling. Figure 1 shows a simple image caption generation process as proposed by Wang et. al. [1].

Image captioning has multiple purposes, e.g., aiding visually impaired individuals in understanding image content, improving social media platforms by enhancing captions on images, and assisting children in learning about objects through images.

Extracting meaningful captions in natural language (such as English) from images, which involves identifying various objects in the image, is considered a challenging and complex. It is a remarkable feat of artificial intelligence to emulate the human ability to caption pictures.

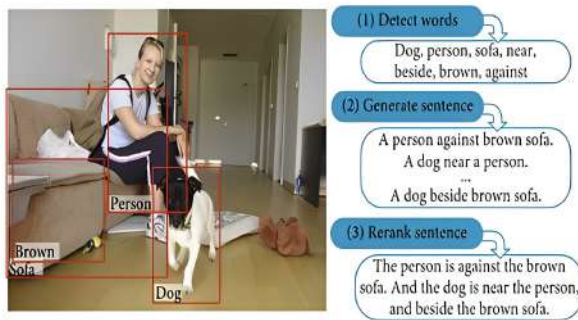


Fig. 1: Shows a sample image caption generation process as proposed by Wang et. al.[1]. This process is divided into three stages, word detection, sentence generation, and sentence re-ranking.

Originally, computer systems relied on predefined formats to generate text descriptions for images, which often lacked the variety needed for rich textual descriptions. The evolution of artificial neural networks (ANNs) has helped overcome these limitations. An intelligent system's task is to identify, recognize, understand, and describe images with captions. An image caption generation model involves identifying objects in digital images to connect them with natural languages, like English. Recently it was observed through studies that attention-based models are effective in capturing the context of an image and establishing relationships between different features of that image.

Various image captioning techniques have been proposed and have been well-received within scientific communities in the past years. However, these techniques have not fully met the demands of society. Consequently, there has been an evolution of various image caption generation techniques to address this gap. Therefore, a systematic analysis of those methods in this field is crucial to evaluate their advantages, disadvantages, and performance. This paper aims to analyse and study different methodologies, focusing on points like algorithms and applications along with their strengths and weaknesses.

There are different sections in this study to discuss the main topic. Here, we are using some abbreviations like PR or P-Rate for Precision rate, RR or R-Rate for Recall Rate, for F1 Score FS and A-Score or AS for Accuracy score. Here, section 2 contains discussion of related works in image captioning from previous decades, highlighting the strengths and weaknesses of various methods. The significance of different image caption generation datasets, and various criteria for measuring the performance of these methods. The next part provides an in-depth classification of image captioning, presenting algorithms, equations, and block diagrams. Section 4 discusses the results obtained for different parameters using various methods described in the 'Related Works' section. These results are presented using tables and graphs to determine the superior method in terms of various performance parameters.

2 Related Works

Various caption-generation techniques from images have been proposed during the last two decades.

Attentive linear transformation[2] operates on linear transformation weights, encoding valuable information in an abstract manner. This technique has shown high performance, enabling the model to generate answers to questions based on images and predicts missing words to complete sentences. However, it has limitations in identifying words associated with sign language, segregating intersecting objects, calculating quantities accurately, and once in a while determining the gender of a human. A Multimodal Method[3] has been proposed, incorporating RNN and LSTM attention-based methods for image caption generation, specifically on the RSICD dataset. While capable of generating captions, this method may produce sentences that overwrite previous ones in remote-sensing images, limiting its uniqueness. This method can not be considered unique for image caption generation. A Topic-oriented model[4] has demonstrated satisfactory results on the MSCOCO dataset, generating high-quality captions. However, its application has not been extended to other datasets. The Multitask Learning Algorithm for Cross-Domain Image Captioning [5] achieves high performance but struggles to distinguish images with visual symmetry and provides low scores for random samples. The Context Sequence Memory Network (CSMN) [6] is adept at extracting various context types from query images, storing long-term information, and understanding context. However, it lacks the inclusion of various metadata and has limitations in social media applications, suggesting room for improvement in post commenting. The Unsupervised Cross-Media Alignment[7] model performs alignment of phrases, and word text conversion, and supports multiple languages but suffers from accuracy issues. Stack-VS [8] is a stack decoder-based model that generates visual and semantic level captions, and fine-grained captions. It cannot consistently generate reasonable captions, suggesting the potential incorporation of a GCN model for enhanced results. A Multimodal attribute detector and subsequent attribute predictor [9] model excels in dynamic attribute prediction and precise caption generation but falls short in producing relational attributes. The Visual Attention Model [10], a cross-lingual model with an independent recurrent structure, conducts feature and semantic similarity analysis but struggles with language-specific tasks and information loss. The multi-level policy and reward RL framework [11] focuses on word and sentence-level caption generation, vision-language, and language-language reward. It requires training the policy network to generate output. The space-time-based memory attention [12] model establishes a strong temporal connection between attention and performance, learning the space and time-based relationship of attended areas. However, this model in no way is a common one. The global-local discriminative objective [13] model enhances discriminability and is proficient in image caption generation and fine-grained caption generation. However, adjusting value of the local discriminative threshold might be needed, as the current setup tends to

overwrite frequently used words. Visual Semantic Attention Model (VSAM) [14] visual keyword concept generates precise and valuable captions and it is effective for visual keyword extraction. This image captioning framework is not fully-developed. The precision rate of this model needed to be improved. Adversarial reinforced report-generation framework [15] is a novel X-ray caption generation framework that is not capable of removing noise from the X-ray images. Bidirectional depth residuals gated recurrent unit network [16] gives a high prediction rate and low inference time. However, this model has poor stability. A Noise Augmented Double-stream Graph Convolutional Networks (NADGCN) [17] model is capable of extracting full image context. Even background context is extract-able using this model but it has limited versatility. This model's drawback is the architectural complexity. The NICVATP2L [18] model achieves low accuracy in language generation and exhibits limited diversity in multi-attribute entities, yet it can produce descriptive and informative captions. Semantic-Constrained Self-learning (SCS) [19] model excels in effective semantic object detection and delivers cutting-edge unpaired captioning. Due to high cost and need of complex experimental setup however, this model is not generic. Context-Aware Visual Policy network (CAVP) [20] model is capable of efficient sentence captioning and produces improved paragraph captioning. However the improvement of sentence, paragraph captioning, and performance matrices needed to be observed, and also improvement for sequential decision-making operations can be done. The task-adaptive Attention module [21] is a non-visual features extraction technique and expression to caption conversion. And potential for improved performance can be there for this model. This method could be applied to attention-based encoder-decoder image captioning. Context-Driven Extractive Method [22] is capable of performing a good estimation of the context from multiple sources, but it is not an effective way of finding annotation. Visual-Semantic Alignment [23] is capable of generating a description using one input array, but it has a significant drawback in producing region-level captions. Gradual Transition and Scope-Caption Detection [24] is a computationally efficient robust methodology. But the application on any large dataset was never demonstrated. An Attention Mechanism [25] is proposed for generating and managing image captions, but it lacks a resultant table and requires numerous performance metrics to determine the final results. The ECANN Model [26] proposes to capture the image captioning from the Grocery datasets, which is a combination of two different datasets. The model is a combination of CNN and LSTM. To optimize the loss during caption generation, the AAS optimization technique is used. The performance of this method is measured using various performance parameters. The model produces high accuracy, efficiency, and performance but the model is incapable of producing captions for multiple languages. The model is not applied to mobile apps. The model is not embedded with NLP. A combination of WCNN, VAPN, and LSTM is proposed. The WCNN model extracts spatial-visual features, while the VAPN model focuses on attentive-visual features

of the image. To calculate the probability of correct word prediction LSTM Module is used. The Model uses MSCOCO and Flickr datasets and produces the results in terms of the CIDr parameter. The model proved to be effective in terms of other existing methods but failed to produce the finer caption.

Table 1 presents various methodologies proposed by different authors over the years, along with the datasets used and performance parameters.

Table 1: Comparison Table I

Year	Author(s)	Methodology	Dataset(s)	Performance Metrics
2010	P. Pham et.al.	UCMA Model	Labeled Faces in the Wild	F1-Score and Others
2015	S. Ye et. al.	ALT Model	MS-COCO + Flickr	BELU and Others
2015	C. C. Park et. al.	GSMN Model	Insta PIC + YFCC	BELU and Others
2015	C. Yan et. al.	TAAAM Model	MS-COCO	BELU and Others
2015	A. Karpathy et. al.	VSA Model	Flickr + MS-COCO	BELU and Others
2016	J. Ali et.al.	GT+SCD Model	Sports Video	Accuracy Rate
2017	X. Lu et. al.	MM Framework	RSICD	BELU and Others
2017	A. Tariq et. al.	GDE Method	News Image	Mean Precision and Others
2018	N. Yu et. al.	TO Model	MS-COCO + Flickr	BELU and Others
2018	N. Xu et. al.	MLP+RRL Framework	Flickr + MS-COCO	BELU and Others
2019	Z. J. Zha et.al.	CAVP Framework	MS-COCO + SIPC	BELU and Others
2020	M. Yang et. al.	MLADIC Framework	MS-COCO + Flickr + Oxford-102	BELU and Others
2020	L. Cheng et. al.	Stack-VS Model	MS-COCO	BELU and Others
2020	Y. Huang et. al.	MAD+SAP Model	MS-COCO	BELU and Others
2020	B. Wang et. al.	VA Model	Flickr	BELU and Others
2020	J. Ji et. al.	STMA Model	MS-COCO	BELU and Others
2020	J. Wu et. al.	GLDO Model	MS-COCO	BELU and Others
2020	L. Wu et. al.	NADS+GCN Framework	MS-COCO	BELU and Others
2020	M. Liu et. al.	NICVATP2L Model	Chinese AIC-ICC	BELU and Others
2020	H. Yanagimoto et. al.	Proposed Attention Model	MS-COCO	Cross-Entropy Loss
2021	S. Zhang et. al.	VSAM Model (VSAM)	IVKD	Precision and Others
2021	D. Hou et. al.	AR+RG Framework	IU X-Ray + MIMIC-CXR	BELU and Others
2021	Z. Zhou et. al.	BDR+GRN Model	MS-COCO	BELU and Others
2021	H. Ben et. al.	SCS Model	Flickr + MS-COCO	BELU and Others
2022	T. Tiwary et. al.	ECANN Model	GSD + FGD	Accuracy and Others
2023	R. Sasibhooshan et. al.	WCNN+VAPN+LSTM Model	Flickr + MSCOCO	CIDEr

3 Methodologies

This part covers discussion of various image caption analysis methodologies. Detailed classification of all these methodologies is presented using Fig. 2. We classify all the methodologies into two sections. The first method is the primitive approach, which is divided into three sections: generic, search-based, and language template-based. Second, is the deep neural network-based methods which are further classified into ten categories including attention-based, and transformer-based. The attention-based methods may be soft, hard, multi-head, or others, depending on their design and construction. In subsection 3.1, we have explained the brief methodology of all the methods demonstrated in Figure. 2.

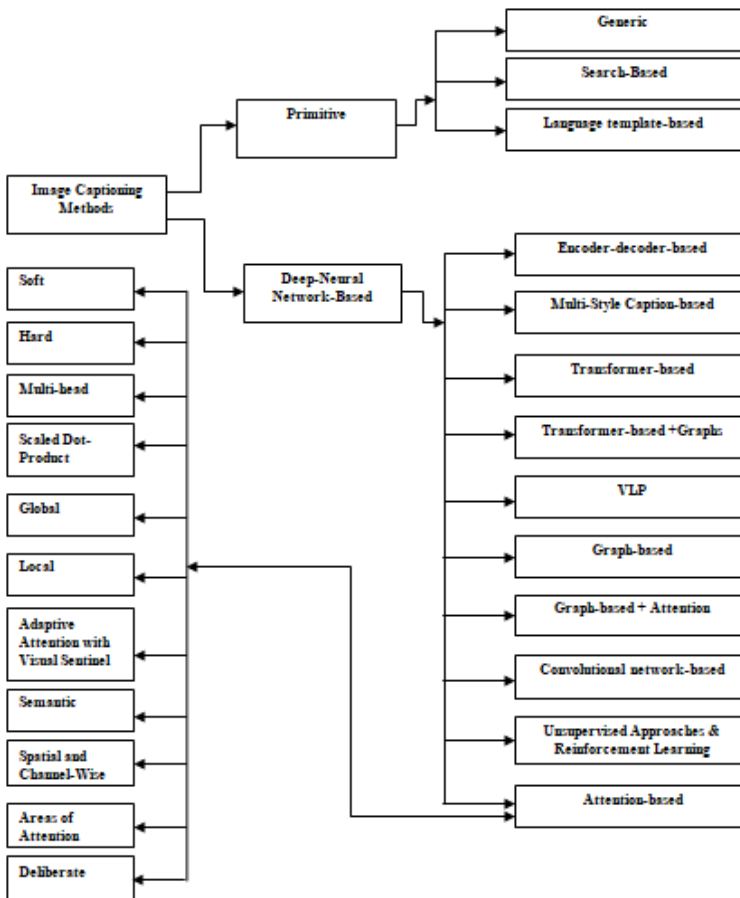


Fig. 2: Shows various methodologies related to the image caption generation as proposed in [27]

3.1 Primitive Methodologies

The primitive method [28] is segregated into three categories such as generic, search-based, and language-template-based. The mathematical, algorithmic, or block diagram-based explanation of these categories is mentioned in subsections 3.1.1, 3.1.2, and 3.1.3 respectively.

3.1.1 Generic Methodology

In detail the primary goal is outlined in Algorithm number 1. The primary goal is to generate text captions from input images using a straightforward, pipeline-based automatic approach.

Algorithm 1 Generic Methodology

Require: Query Image $I(p, q)$

Ensure: Text Caption

```

1: while number of image in the dataset  $\neq 0$  do
2:   Consider a query image  $I(p, q)$  from the dataset
3:   Extract all the information present in  $I(p, q)$ 
4:   Find the textual content,  $T_{pq}$ , from  $I(p, q)$ 
5:   Find the visual content,  $V_{pq}$ , from  $I(p, q)$ 
6:   Establish a connection between the  $T_{pq}$  and  $V_{pq}$ 
7:   Ensure the language fluidity present in the  $I(p, q)$ 
8:   if Steps 3 and 4 are guaranteed then
9:     Converts the information present in the  $I(p, q)$  into text caption
10:    Print the output text caption
11:   else
12:     Print Error
13:   end if
14:   Find the next suitable  $I(p, q)$  from the dataset
15: end while

```

3.1.2 Search-Based Methodology

The primary objective of the search-based methodology [28] is to generate the image caption (text) from the given query image from the dataset. Several researchers propose various versions of search-based image caption generation methods. These methods suffer from, heavy dependency on the dataset, incapability of producing new sentences, very few high-quality existing training captions in the training dataset to produce the final caption, various content and style differences between the query image and training image, and unsatisfactory output caption. The specific steps of this approach are described in detail in Algorithm 2.

Algorithm 2 Search-Based Methodology**Require:** Query Image (I)**Ensure:** Text Caption (T)

- 1: **while** *number of image in the dataset* $\neq 0$ **do**
- 2: Consider a query image I from the dataset
- 3: Consider the training dataset, T_d
- 4: Construct the searching image, S using I
- 5: Compare I and S
- 6: Search identical images from T_d using Step 4
- 7: Find the caption present in the identical images from T_d
- 8: Marked all the text captions as a candidate
- 9: Re-rank all text captions
- 10: Determine a detailed image description
- 11: Print the output text caption
- 12: Find the next suitable I from the dataset
- 13: **end while**

The methodology's block diagram is elucidated in Figure. 3 as follows:

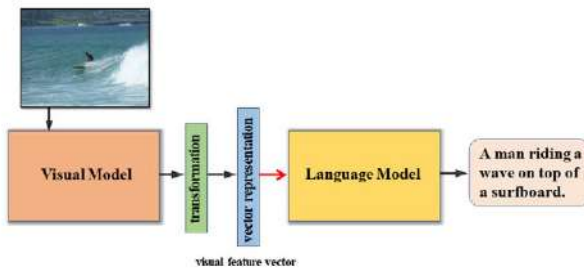


Fig. 3: Shows Search-Based Methodology as proposed in [28]

3.1.3 Language Template-Based Methodology

A basic understanding of the query image's visual feature is an absolute requirement in this approach. This step is followed by the visual feature extraction strategies. Finally, the descriptive caption is generated using a language template model. In comparison with other existing models, this model proves to be a better visual content generator. Due to the inclusion of a language template, this model is capable of producing grammatically correct sentences. Over the years several researchers have proposed several language template-based methodologies. Simplicity, homogeneity, artificiality, and inarticulate descriptive caption make these models very weak. Algorithm 3 describes the detailed procedure of this approach, whereas, Figure. 3 describes a language template-based [28] image captioning approach.

Algorithm 3 Language Template-Based Methodology

Require: Query Image; $I(p, q)$, Transformation Function; $f(V)$, Visual Feature Vector; \vec{V}

Ensure: Descriptive Captions

- 1: **while** *number of image in the dataset* $\neq 0$ **do**
- 2: Consider a query image $I(p, q)$ from the dataset
- 3: Construct a visual Model, V_{model}
- 4: Extract the visual contents, from $I(p, q)$ using V_{model}
- 5: Transform visual contents values to \vec{V} using $f(V)$
- 6: Pass \vec{V} through the language template model, LT_{model}
- 7: Produce descriptive captions through LT_{model}
- 8: Print descriptive captions
- 9: Find the next suitable $I(p, q)$ from the dataset
- 10: **end while**

The methodology's block diagram is elucidated in Figure. 4 as follows:

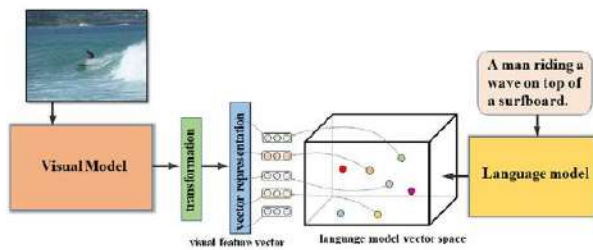


Fig. 4: Shows a combined framework of Search-based and language template-based Methodology as proposed in [28]

3.2 Deep Neural-Network-Based Methodologies

With the evolution of the deep neural network, deep-neural-network-based image caption generation methods [28] have evolved in recent years. These methods take query images as input. To extract visual information from an image, a visual model is employed, constructed using deep neural network-based architectures like attention-guided CNNs, channel-attention-based CNNs, self-attention-based models, GCNs, and others. This model identifies and captures attributes, objects, and other visual elements present in the image. Similarly, a language model is constructed using one or more deep neural network-based architectures such as CNN, CNN+RNN, etc. The visual model generates the visual features specific to the query image as its output. These features are represented with the help of visual feature vectors, which are transmitted into language models later on. Both models are trained using

a range of strategies, including masked language model training, cross-entropy loss, VL pre-training, reinforcement learning, etc. Finally, the text caption is received as the output. The model ultimately produces a text caption as its output. A block diagram of this methodology is shown in Figure 5.

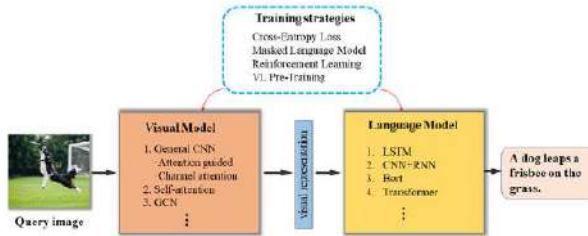


Fig. 5: Shows a sample Deep Neural-Network-Based Methodology as proposed in [28]

3.2.1 Generic Deep-Neural-Network-Based Methodology

There are many preexisting methods related to Generic Deep-Neural-Network-Based Methodology [29], but a generic architecture [30] of this model is considered here to demonstrate its working principle as described in Algorithm. 4 and Figure. 6.

Algorithm 4 Generic Deep-Neural-Network-Based Methodology

Require: Query Image

Ensure: Text Captions

- 1: An image input is given to the system.
 - 2: Then the input is processed by CNNs before passing it to the next level.
 - 3: The subsequent stage consists of components for extracting visual features, information of any celebrity or landmark, and feature vectors. This feature extraction model then gets processed by a deep Residual Neural Network model.
 - 4: The results from the VCE, CIE, and LIE are inputted into the language-based model.
 - 5: The output from the feature vector separator is fed into a DMSM.
 - 6: The language and DMSM-based model's output is lodged into the Confidence Generation Model.
 - 7: The output caption is generated by the Confidence Generation Model.
-

The methodology's block diagram is elucidated in Figure. 6 as follows:

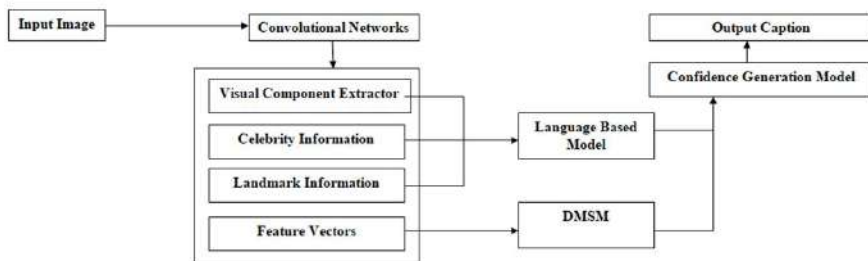


Fig. 6: Shows a Generic Deep Neural-Network-Based Methodology as proposed in [29]

The proposed algorithm suffers from poor robustness, limitations towards one or few datasets, poor testing results (as very few performance parameters are tested), and low-performance accuracy.

3.2.2 Encoder-Decoder Based Methodology

There are many preexisting methods related to the Encoder-Decoder-Based Methodology [31], but a generic architecture of this model is considered here to demonstrate its working principle as described in Algorithm. 5 and Figure. 7.

Algorithm 5 Encoder-Decoder Based Methodology

Require: Query Image

Ensure: Text Captions

- 1: The feature vector of a query image (P) is fed into an LSTM sequence with a time interval of -1 .
 - 2: The entire word sequence is taken as input from time interval $t = 0$ and onwards.
 - 3: The highest probability at each step is calculated using the hidden state's activation function.
 - 4: Then Beam Search is employed to accurately predict captions of different images. At each time step t , it evaluates m sentences as the best candidates for prediction. For the subsequent time step $t + 1$, it considers the most likely words from these m sentences, resulting in m^2 potential sentences. From these, the best probable sentences, m are selected for the next step.
 - 5: The endmost output, whether a word or a text caption, is produced through the application of Equations 1, 2, and 3.
-

The methodology's block diagram is elucidated in Figure. 7 as follows:

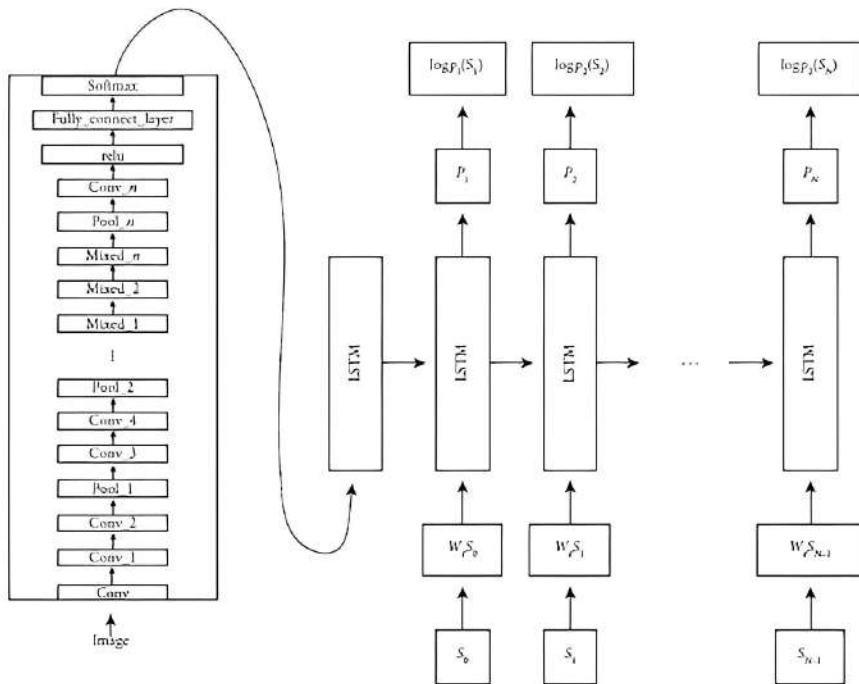


Fig. 7: Shows a Encoder-Decoder Based Methodology a proposed in [31]

The encoder-decoder model suffers from many disadvantages. If RNN is used for the decoder model the model suffers from high model training cost, global gradient descent problem, and remembering the contents of the previous iterations.

3.2.3 Attention-Based Methodology

After receiving information from various resources, human beings consciously ignore some of the major and minor information. This self-selection ability of human beings is known as the attention mechanism [32]. These methods are inspired by the way the human being observes an image. The method focuses on two major points: (a) Which part of the query image needs to be focused? (b) how to allocate information-extracting resources to the parts of the query image identified in step (a). The attention method is implemented using two steps. In the first step, the target module t_m and the source module s_m are linked using $f(t_m, s_m)$ function as demonstrated in Eq. 4. In the second step, α_t or the final text caption is generated using the probability distribution function as demonstrated in Eq. 5.

The model is segregated into twelve categories including soft-attention-based, hard-attention-based. The mathematical, algorithmic, or block diagram-based explanation of these categories is mentioned in the paragraphs given below.

Soft-Attention-Based Methodology

There are many preexisting methods related to Soft-Attention-Based Methodology [33][34], but a generic architecture of this model is considered here to demonstrate its working principle as described in Algorithm. 6 and Figure. 8.

Algorithm 6 Soft-Attention-Based Methodology

Require: Query Image; $I(p, q)$, i/p Sequence; S_t , a word in S_t ; w , Context vector; \vec{C}_v

Ensure: Text Captions

- 1: **while** *number of image in the dataset* $\neq 0$ **do**
 - 2: Consider a query image $I(p, q)$ from the dataset
 - 3: Create an LSTM model
 - 4: Calculate the weighted image features(Z) using the following steps:
 - 5: Divide $I(p, q)$ into various subsections such as, x_1, x_2, x_3 , and x_4
 - 6: For each section x_i calculate a score s_i using Eq. 6
 - 7: Use s_i for normalization
 - 8: α_i weight is quantify using Eq. 7
 - 9: Quantify the average weight of x_i using Eq. 8
 - 10: The LSTM input x is restored by Z
 - 11: Produce text captions
 - 12: Print text captions
 - 13: Find the next suitable $I(p, q)$ from the dataset
 - 14: **end while**
-

The equations used in Algorithm 6 are calculated as follows:

$$\varepsilon_{p(\alpha_t|\beta)} [\vec{C}_v] = \sum_{i=1}^k \delta_{t,i} \cdot \beta_i \quad (1)$$

$$\theta(\beta_i, \delta_i) = \sum_k^i \delta_{t,i} \cdot \beta_i \quad (2)$$

$$k = -\log f(q | p) + \lambda \cdot \sum_i^k (1 - \sum_i^k \delta_{t,i}) \quad (3)$$

The methodology's block diagram is elucidated in Figure. 8 as follows:

Due to the parameterized approach, the soft attention model can be used for direct training, which is considered an advantage of this model. One of

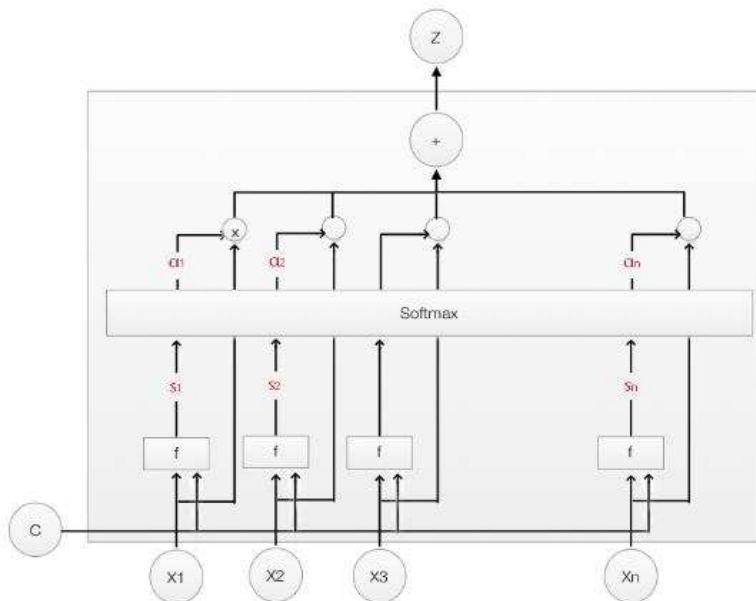


Fig. 8: Shows a Soft-Attention-Based Methodology as proposed in [34]

the major disadvantages of the model is training time. If we add multiple parameters to measure the weights then the model becomes time-consuming.

Hard-Attention-Based Methodology

There are many preexisting methods related to Hard-Attention-Based Methodology [33][34], but a generic architecture of this model is considered here to demonstrate its working principle as described in the Algorithm. 7 and Figure. 9.

Algorithm 7 Hard-Attention-Based Methodology

Require: Query Image (I)**Ensure:** Text Captions (T)

```

1: while number of image in the dataset  $\neq 0$  do
2:   Consider a query image  $I$  from the dataset
3:   Create an LSTM model
4:   Calculate the sample rate ( $SR_{rate}$ ) using the following steps:
5:   Divide  $I$  into various subsections such as,  $y_1, y_2, y_3,$  and  $y_4$ 
6:   For each section  $y_i$  calculate a score  $sr_i$  using Eq. 6
7:   Use  $sr_i$  for normalization
8:    $a_i$  weight is quantify using Eq. 7
9:   Quantify the average weight of  $y_i$  using Eq. 8
10:  Use The LSTM input  $y$  is restored by  $sr_{rate}$ 
11:  Produce  $T$ 
12:  Print  $T$ 
13:  Find the next suitable  $I$  from the dataset
14: end while

```

The equations used in Algorithm 7 are calculated as follows:

$$sr_i = htf(W1_{c1} \cdot C1 + W1_a \cdot A_i) = htf(W1_{c1} \cdot h1_{t1-1} + W1_a \cdot y1_i) \quad (4)$$

$$a_i = sftmf(sr_1, sr_2, sr_3, \dots, sr_i, \dots) \quad (5)$$

$$SR_{rate} \sim a_i, a_i \quad (6)$$

One of the disadvantages of this model is the incapability of calculating gradient back-propagation on its own. The model needs Monte Carlo sampling in this regard. The model was unable to establish any relation of attention distribution and the final error (loss) function. The back-propagation algorithm also fails to achieve enough training.

This methodology's block diagram is illustrated in figure below:

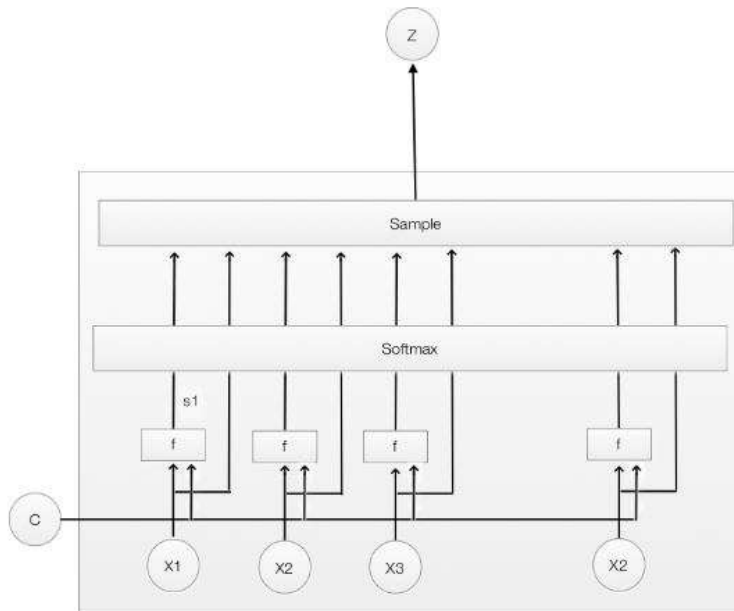


Fig. 9: Shows a Hard-Attention-Based Methodology as proposed in [34]

Multi-Head-Attention-Based Methodology

There are many preexisting methods related to Multi-Head-Attention-Based Methodology [33], but a generic architecture of this model is considered here to demonstrate its working principle as described in Algorithm. 8 and Figure. 10.

Algorithm 8 Multi-Head-Attention-Based Methodology

Require: Query Image; $I(p, q)$

Ensure: Text Captions

- 1: **while** *number of image in the dataset* $\neq 0$ **do**
 - 2: Consider a query image $I(p, q)$ from the dataset
 - 3: Extract the q , k , and v information from $I(p, q)$
 - 4: Create 3 layers for the q , k , and v
 - 5: Create weight for each layer
 - 6: Create the q , k , and v matrices
 - 7: Segregate the matrices over the considerable heads of attention
 - 8: Partition input data using logical split and uniform distribution of linear layer's weight over the considerable heads of attention.
 - 9: Choose size of q and compute it using Eq. 12
 - 10: Reshape linear layers and incorporate head dimension using Eq. 13
 - 11: Generate the output of q , k , and v matrices
 - 12: Compute the attention score for each head using Eq. 13
 - 13: Each attention head's scores are clubbed into a single score
 - 14: Produce text captions
 - 15: Print text captions
 - 16: Find the next suitable $I(p, q)$ from the dataset
 - 17: **end while**
-

The equations used in Algorithm 8 are calculated as follows:

$$\text{Query Size} = \frac{\text{Embedding Size}}{\text{Number of heads}} \quad (7)$$

$$\text{multiple}f(q, k, v) = \text{concatenationfunction}(h1_1, h1_2, \dots, h1_h) \cdot w^o, \quad (8)$$

Where, $h1_i = \text{attentionfunction}(q \cdot w_i^q, k \cdot w_i^k, v \cdot w_i^v)$

$$\text{attention score}(a_s) = \text{softmaxfunction} \left(\frac{q * k^t + \text{mask}}{\sqrt{\text{Query Size}}} \right) * v \quad (9)$$

The methodology's block diagram is elucidated in Figure. 10 as follows:

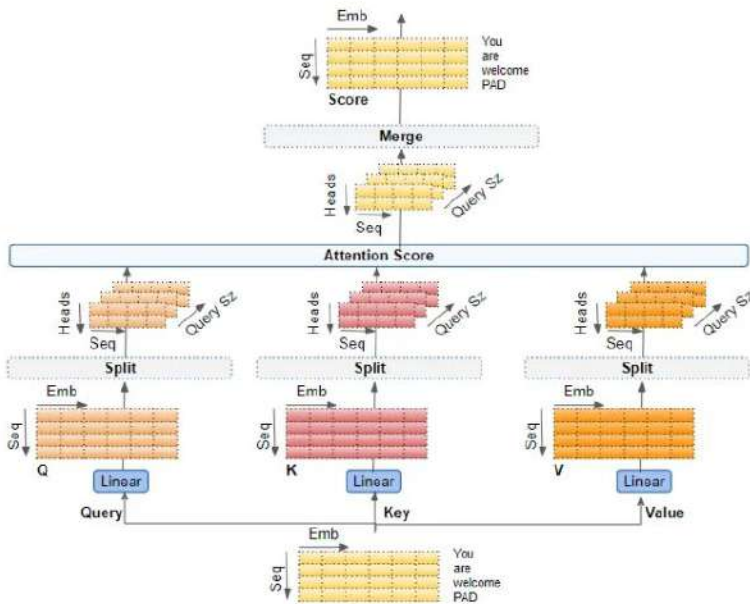


Fig. 10: Shows a Multi-Head-Attention-Based Methodology as proposed in [35]

One of the major disadvantages of the model is the low-rank bottleneck, which means that the rank of the attention weight matrix is too small to represent any desired attention.

Scaled Dot-Product Attention-Based Methodology

There are many preexisting methods related to Scaled Dot-Product Attention-Based Methodology [33], but a generic architecture of this model is considered here to demonstrate its working principle as described in Algorithm. 9 and Figure. 11.

The disadvantage of this model is the vanishing gradient problem, poor dot-product attention calculation due to large dimension, and high complexity due to highly optimized matrix multiplication code.

Algorithm 9 Scaled Dot-Product Attention-Based Methodology

Require: Query Image; $I(p, q)$

Ensure: Text Captions

```

1: while number of image in the dataset  $\neq 0$  do
2:   Consider a query image  $I(p, q)$  from the dataset
3:   Extract the  $q$ ,  $k$ , and  $v$  information from  $I(p, q)$ 
4:   Quantify dot-products of  $q$  with the help of  $k$ 
5:   The result of Step 3 is scaled by  $d_k$ 
6:   Produce the attention score
7:   Attention scores are fed into a softmax function
8:   Obtain a set of attention weights using Eq. 15
9:   Attention weights are used to scale the values through a weighted
   multiplication operation
10:  Produce text captions
11:  Print text captions
12:  Find the next suitable  $I(p, q)$  from the dataset
13: end while

```

The calculation for the equation utilized in Algorithm 9 is done in this way:

$$attwgt(q, k, v) = softmax\left(\frac{q \cdot k^t}{\sqrt{d_k}}\right) \cdot v \quad (10)$$

The methodology's block diagram is elucidated in Figure. 11 as follows:

Global Attention-Based Methodology

There are many preexisting methods related to Global Attention-Based Methodology [36], but a generic architecture of this model is considered here to demonstrate its working principle as described in the Algorithm. 10 and Figure. 12.

The model decides to concentrate on all inputs of the encoder while computing each decoder state. This leads to huge calculating steps. This is considered one of the major disadvantages of this model.

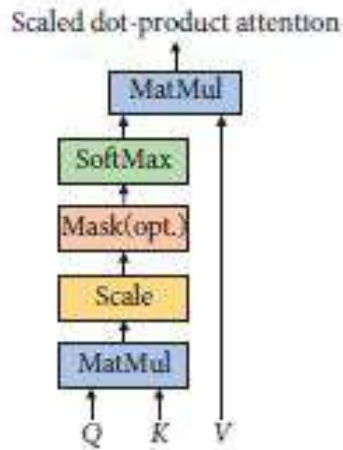


Fig. 11: Shows a Multi-Head-Attention-Based Methodology as proposed in [28]

The methodology's block diagram is elucidated in Figure. 12 as follows:

Local Attention-Based Methodology

Local Attention [36] is an attention mechanism that focuses on a limited set of hidden states when forming the context vector. This subset can be determined through techniques such as Monotonic Alignment and Predictive Alignment. In the Monotonic technique, the subset of hidden states is chosen by retaining those that are closer to the current translation step. Conversely, the Predictive technique selects hidden states in proximity to a predicted position, factoring in the word being translated at that moment. The primary benefit of local attention lies in its ability to lower computational burdens on the attention mechanism. While there exist numerous preexisting methods related to Local Attention-Based Methodology, this discussion centres around a general model architecture to illustrate its operational concept, as outlined in Algorithm 11 and depicted in Figure 13.

Algorithm 10 Global Attention-Based Methodology

Require: Input Sequence (I)**Ensure:** Text Captions

- 1: **while** *number of image in the dataset* $\neq 0$ **do**
 - 2: Provide I to the encoder (X)
 - 3: X encodes I and generates the output (hs)
 - 4: Decoder (Y) computes hs and produces target decoding (ht)
 - 5: Use ht for generating each encoded time step score, and use the softmax function for normalizing each score. Use the following scoring functions to execute step 5:
 - 6: dot: Perform dot-product between ht and source encoding (se)
 - 7: general: Perform dot-product between ht and weighted se (wse)
 - 8: concat: Perform concatenation between se and ht
 - 9: location: a softmax of the weighted target decoding
 - 10: Create Context Vector (V) with the help of the weighted sum (ws) of ws and the alignment weight (aw)
 - 11: Calculate V concat ht , V weighted ht , V transferred ht using a tanh function
 - 12: Produce text captions
 - 13: Print text captions
 - 14: Find the next suitable I from the dataset
 - 15: **end while**
-

Algorithm 11 Local Attention-Based Methodology

Require: Input Sequence, I **Ensure:** Text Caption, T

- 1: **while** *number of image in the dataset* $\neq 0$ **do**
 - 2: Insert I to the encoder(X)
 - 3: Search an alignment position (ap)
 - 4: Computes the left attention weight (law)
 - 5: Computes the right attention weight (raw)
 - 6: Measure the weight of the context vector (CV)
 - 7: Produce T
 - 8: Print T
 - 9: Find the next suitable I from the dataset
 - 10: **end while**
-

The methodology's block diagram is elucidated in Figure. 13 as follows:

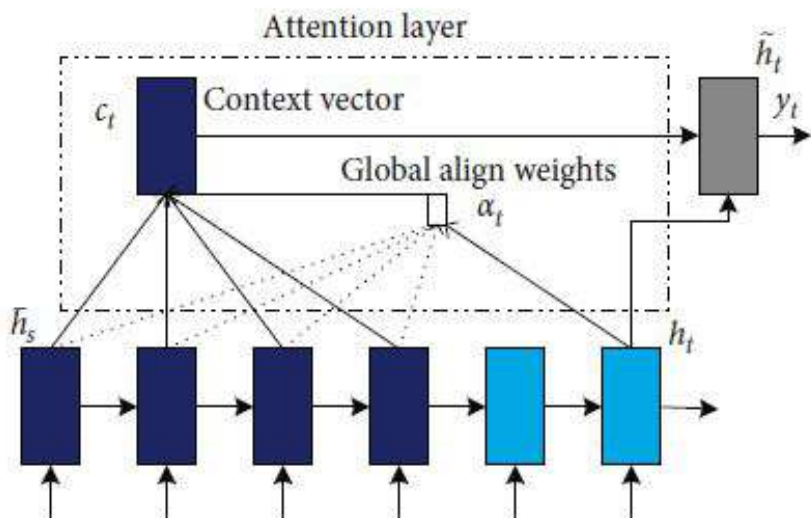


Fig. 12: Shows a Multi-Head-Attention-Based Methodology as proposed by [28]

Adaptive Attention with Visual Sentinel-Based Methodology

There are many preexisting methods related to Adaptive Attention with Visual Sentinel-Based Methodology [37], but a generic architecture of this model is considered here to demonstrate its working principle as described in Algorithm.12 and Figure. 14.

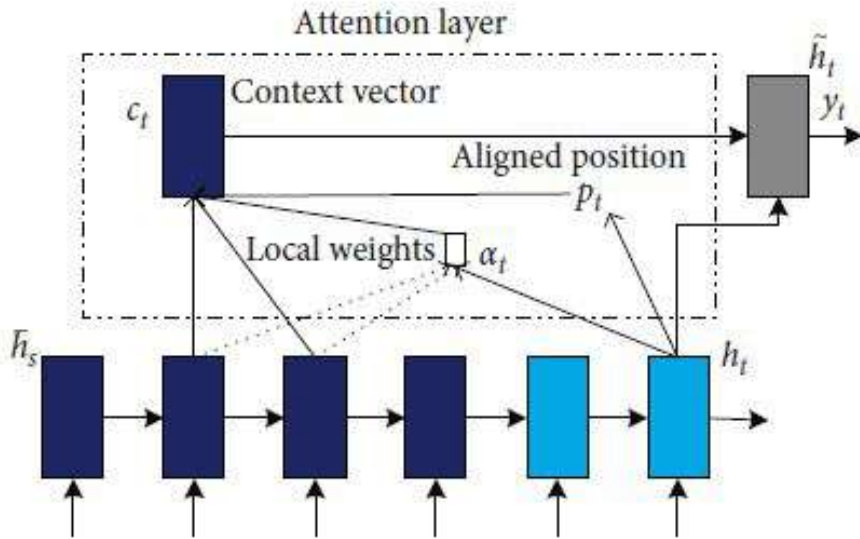


Fig. 13: Shows Local-Attention-Based Methodology as proposed by [28]

The methodology's block diagram is elucidated in Figure. 14 as follows:

The method is capable of determining the amount of new information from the query image during the network training procedure. This is considered the advantage of Adaptive Attention with Visual Sentinel-Based Methodology. The model is incapable of distinguishing non-visual and visual words with high correlation, due to un-reliability on visual signal. This is considered a disadvantage of this model.

Algorithm 12 Adaptive Attention with Visual Sentinel-Based Methodology**Require:** Query Image; I **Ensure:** Text Captions

- 1: Create an LSTM model using the following sub-steps: Create the forget gate as follows: $fgt \leftarrow s \cdot (\Omega_\zeta \cdot [\eta_{\tau-1}, \chi_\tau] + \beta_\zeta)$
- 2: Create the input gate i , which is divided by two branches, ι_τ and $\tilde{\Gamma}_\tau$ respectively. Both branches are constructed using the following steps:
- 3: Create ι_τ branch to determine which cell is updated and which data is updated.
- 4: This is created as follows: $\iota_\tau \leftarrow s \cdot (\Omega_i \cdot [\eta_{\tau-1}, x_\tau] + \beta_i)$
- 5: Create $\tilde{\Gamma}_\tau$ to determine the amount of data added to a cell as follows: $\tilde{\Gamma}_\tau \leftarrow htf(\Omega_\Gamma \cdot [\eta_{\tau-1}, \chi_\tau] + \beta_\Gamma)$
- 6: Update the context vector, Γ_τ using the following formula: $\Gamma_\tau \leftarrow \zeta_\tau * \Gamma_{\tau-1} + \iota_\tau * \tilde{\Gamma}_\tau$
- 7: Produce the final output, omi_τ using the following formula: $omi_\tau \leftarrow s \cdot (\Omega_{omi} \cdot [\eta_{\tau-1}, \chi_\tau] + \beta_{omi})$
- 8: Produce the hidden state, η_τ using the following formula: $\eta_\tau \leftarrow omi_\tau * htf(\Gamma_\tau)$
- 9: Calculate the spatial attention using the following formulas: $p_\tau = \Omega_\eta^T \cdot htf(\Omega_v \cdot v + (\Omega_\theta \cdot \eta_\tau) \cdot 1^T)$, $a_\tau = sftmaxf(p_\tau)$, and $\gamma_\tau = \sum_\lambda a_{\tau_i} v_{\tau_i}$, where, Υ represents any regions in the query image, 1 represents a vector, Ω_σ represents the learned parameters, a is the attention weight over features in Υ , and $\gamma - \tau$ is the context vector.
- 10: Create the sentinel attention as follows: $\theta_\tau = s \cdot (\Omega_\chi \cdot \chi_\tau + \Omega_\eta \cdot \eta_{\tau-1})$, and $\sigma_\tau = \theta_\tau \odot htf(\Gamma_\tau)$ where, σ_τ stands for sentinel visual
- 11: Calculate the sentinel gate, b_τ as follows: $\hat{a}_\tau = sftmaxf([p_\tau; \Omega_\eta^T \cdot htf(\Omega_v \cdot \Upsilon + (\Omega_\theta \cdot \eta_\tau))])$, where, \hat{a}_τ represents the weighted alpha of concatenated p , and $b_\tau = \hat{a}_\tau[\kappa + 1]$
- 12: Construct context vector using sentinel visual information and spatial visual information, using the following formula: $\tilde{\gamma}_\tau = b_\tau \cdot \sigma_\tau + (1 - b_\tau) \cdot \gamma_\tau$
- 13: Compute image captioning using the following formula: $t^* \leftarrow args\ max_t \cdot \sum_I^\psi \log \pi(\psi | I, t)$, where I is the query image, t is model parameter, and ψ is the text output

Semantic Attention-Based Methodology

There are many preexisting methods related to semantic attention-based methodology [38], but a generic architecture of this model is considered here to demonstrate its working principle as described in Algorithm. 13 and Figure. 15.

The equation used in algorithm 13 is explained as follows:

$$X_i = \beta(b_i, Z_j) \quad (11)$$

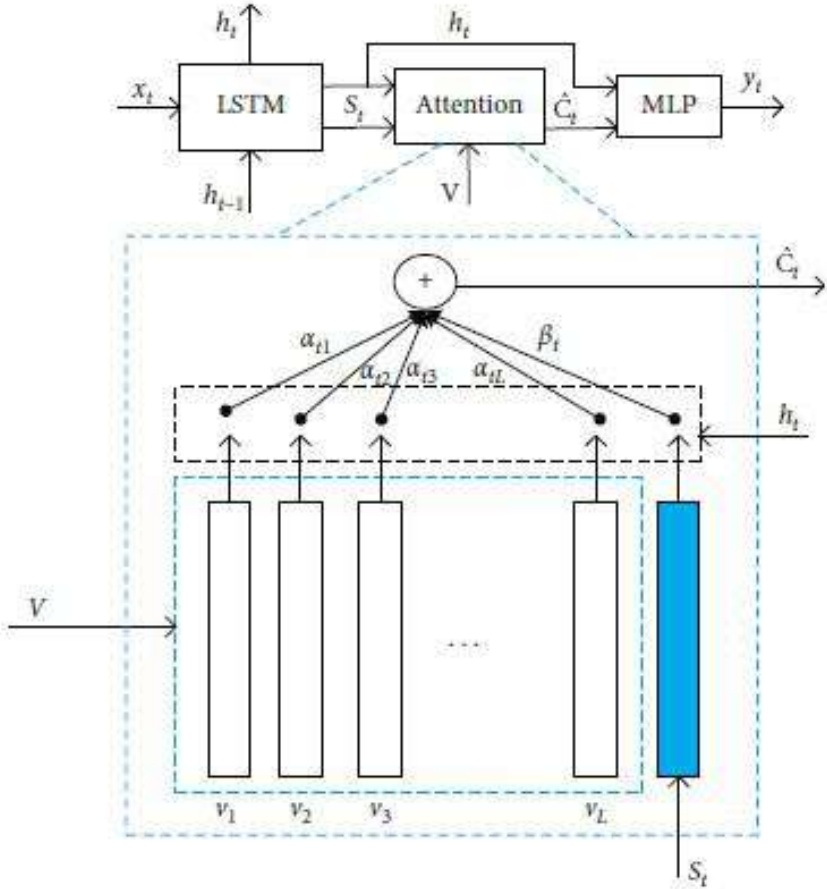


Fig. 14: Shows an Adaptive Attention with Visual Sentinel-Based Methodology as proposed by [28]

where, b_i denotes the i th hidden state input, Z_j denotes the list of visual attributes, $\varphi()$ denotes the output attention model or the output function and X_i denotes the t th output

$$b_i = RNN(b_{i-1}, a_i) \quad (12)$$

where, $RNN()$ denotes the RNN model or operational function, and b_{i-1} denotes $i - 1$ th hidden state input

The Semantic Attention-Based Methodology suffers from several disadvantages. The structure of the model is deep and complex. The model suffers from the overfitting problem, which can be resolved by adjusting and fine-tuning the

Algorithm 13 Semantic Attention-Based Methodology

Require: Query Image; I

Ensure: Y_t

- 1: Consider the query image I
 - 2: From I , extract the visual features (v) using a CNN
 - 3: Calculate the "list of visual attributes", Z_j , using a set of attribute detectors $AttrDet_1, \dots, AttrDet_N$
 - 4: Insert Z_j , and v are fused to get together to generate RNN's output
 - 5: Calculate the value of the initial input node, " a_0 ", using a linear transformation model as follows: $a_0 = \omega^{a,v} \cdot v$, where, $\omega^{a,v}$ denotes the weight of RNN.
 - 6: To traverse existing words and to predict words to be pre-generated in the future, input-output attention models are used.
 - 7: The input model is considered using the following equation: $a_i = \beta(X_{i-1}, Z_j), i > 0$, where, a_i denotes the input to the RNN, $\beta()$ denotes the input attention model or input function, X_{i-1} denotes $i-1$ th output word. X_{i-1} is used as the RNN's feedback input based on the input of a_i
 - 8: Output model is constructed using Eq. 16
 - 9: The b_t is calculated using Eq. 17
-

CNN parameters and using the dropout technique. The value of the loss function may be decreased abruptly for the training set during the model training process.

The methodology's block diagram is elucidated in Figure. 15 as follows:

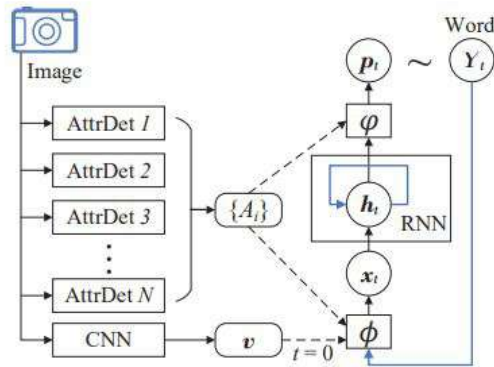


Fig. 15: Shows a Semantic Attention-Based Methodology as proposed in [28]

SCA-CNN-Based Methodology

There are many preexisting methods related to SCA-CNN-based methodology [39], but a generic architecture of this model is considered here to demonstrate its working principle as described in Algorithm. 14 and Figure. 16.

The methodology's block diagram is elucidated in Figure. 16 as follows:

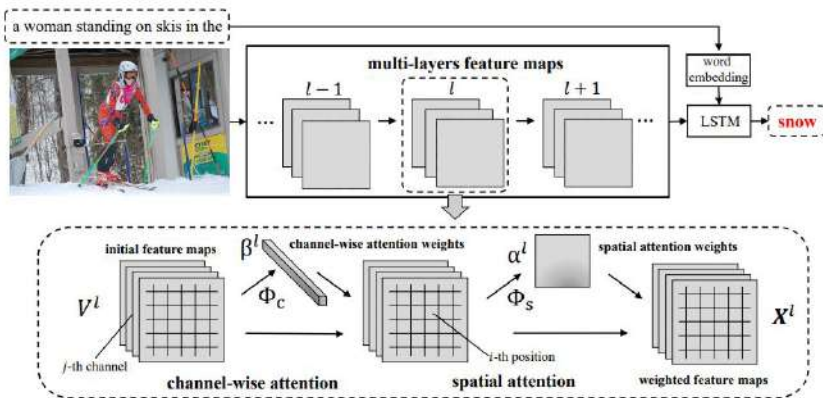


Fig. 16: Shows a SCA-CNN-Based Methodology as proposed by [28]

The method offers attention to be applied in the multiple layers, which is considered to be a useful step to achieve visual attention from the semantic

Algorithm 14 SCA-CNN-Based Methodology

Require: Query Image; I , attention feature map (AFM), channel-wise attention (CWA), channel-wise attention weights (CWA_W), spatial attention (SA), spatial attention weights (SA_W), multi-layer feature map (MLFM)

Ensure: Word (Text representation)

- 1: Consider I
 - 2: Insert I as an input to MLFM layer, $mlfm_i$
 - 3: The $mlfm_i$ works as follows:
 - For the ly th layer, the initial feature map, vt^{ly} is the output of $(ly-1)$ th layer.
 - I is injected to $(ly-1)$ th layer.
 - Output of the previous step is injected to $vt^l vt^{ly}$ layer works as follows:
 - Use CWA procedure, ϕ_{cwa} to obtain CWA_W, β_{ly}
 - Use CWAM to multiply all β_{ly}
 - Use SA procedure, ϕ_{sa} to obtain the α^{ly} or SA_W
 - For each region, multiply all α^{ly} to produce, Z^{ly}
 - Z^{ly} is injected to $ly+1$ th layer of MLFM
 - 4: The final output of MLFM is inserted into the LSTM model
 - 5: A text caption with a missing word is injected into the LSTM layer.
 - 6: Generate missing text through the LSTM layer.
-

abstraction. This is considered to be an advantage of this model. The model suffers from high cost, complexity, and the problem of over-range, which are considered to be the disadvantages of the model.

Areas of Attention-Based Methodology

There are many preexisting methods related to Areas of Attention-based methodology [40], but a generic architecture of this model is considered here to demonstrate its working principle as described in the Algorithm. 15 and Figure. 17.

The model establishes a straight link between the image region and the title word, taking into account the correlation among the state, the image, and the predicted word. This is regarded as an advantage of this method. The relationship between objects and captions is not improved by object training and relation detectors in this approach, which is regarded as a disadvantage of this model.

Algorithm 15 Areas of Attention-Based Methodology**Require:** Query Image (I)**Ensure:** Word (Text representation)

- 1: Consider I
- 2: A CNN model is used to encode I
- 3: Compute vectorized output, $\phi(I)$
- 4: $\phi(I)$, is inserted as an input to a RNN model
- 5: Use $\phi(I)$ to initialize RNN model's current state, W
- 6: Using Eq. 18, create and modify RNN's output, h , iteratively
- 7: Calculate θ_{rh} , θ_{wh} , and θ_{wr} using I and h
- 8: Produce $p(\omega, r)$, using θ_{rh} , θ_{wh} , and θ_{wr}
- 9: Generate word, ω , and pool region descriptor, v , from $p(\omega, r)$
- 10: Insert, ω , and v , as a feedback input to the RNN model.

The equation used in algorithm 14 is explained as follows:

$$h = \theta_{hi} \cdot \phi(I) \quad (13)$$

where, $\phi_{hi} \in \mathbb{R}^{dh \times dI}$

The methodology's block diagram is elucidated in Figure. 17 as follows:

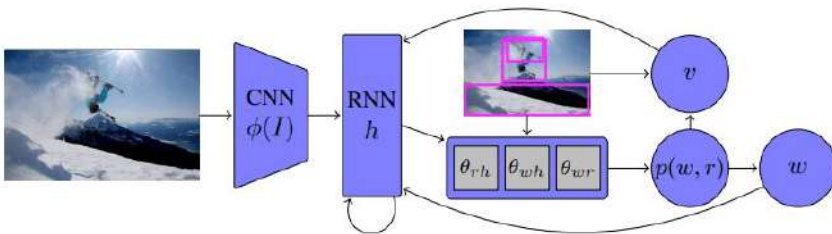


Fig. 17: Shows a Areas of Attention-Based Methodology as proposed by [28]

Deliberate Attention-Based Methodology

There are many preexisting methods related to Deliberate Attention-Based Methodology [41], but a generic architecture of this model is considered here to demonstrate its working principle as described in the Algorithm. 16 and Figure. 18.

The Deliberate attention model fails to generate the captions from those query images from where an individual can engage in reasoning by drawing upon their background knowledge.

Algorithm 16 Deliberate Attention-Based Methodology

Require: Query Image (I), deliberate residual attention network ($DRAN$), residual-based attention layer ($RBAL$), hidden states (HS), visual attention (VA)

Ensure: Word (Text representation)

- 1: Consider I
- 2: Construct a $DRAN$ method that works in two phases
- 3: In phase one, the caption's initial version is generated using $RBAL$'s HS and VA
- 4: In phase two, the caption's initial version is refined using $DRAN$

The methodology's block diagram is elucidated in Figure. 18 as follows:

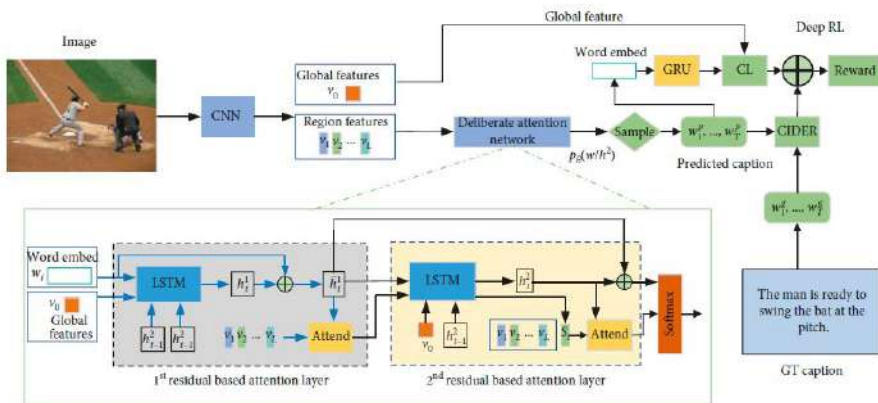


Fig. 18: Shows a Deliberate Attention-Based Methodology as proposed by [28]

3.2.4 Multi-Style-Caption-Based Methodology

There are many preexisting methods related to Multi-Style-Caption-Based Methodology [42], but a generic architecture of this model is considered here to demonstrate its working principle as described in the Algorithm. 17 and Figure. 19.

Algorithm 17 Multi-Style-Caption-Based Methodology**Require:** Query Image; I **Ensure:** Word (Text representation)

- 1: Consider the query image I
- 2: Extract the multi-modality image features using ResNeXt and dense caption neural network
- 3: Construct a multi-up-down fusion network using ResNeXt feature encoder, dense caption feature encoder, previous hidden state top-down decoder
- 4: Feed the multi-modality features into multi-up-down fusion
- 5: Consider the previous word of the text/caption extracted from the query image and construct the embedding matrix
- 6: Consider the personality and construct the embedding matrix constructed from the previous word and personality
- 7: Construct the multi-style component with the help of an embedding matrix constructed from the previous word and personality
- 8: Output of the multi-style component is inserted into the top-down decoder which produces the current hidden states and final word/current word using the linear softmax function

The methodology's block diagram is elucidated in Figure. 19 as follows:

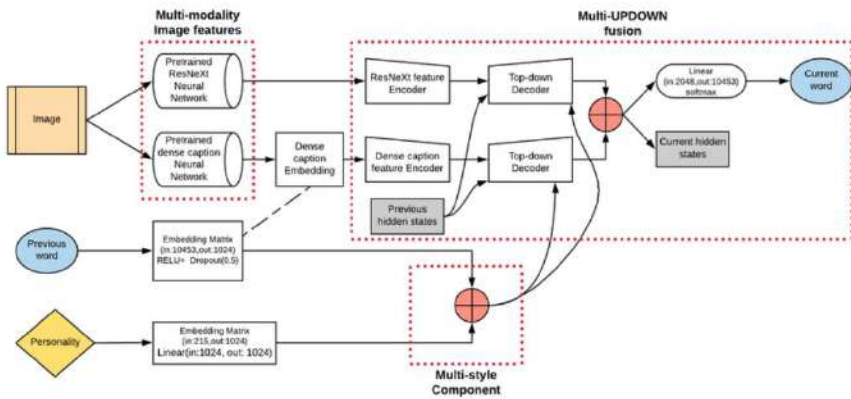


Fig. 19: Shows a Multi-Style-Caption-Based Methodology as proposed by [28]

In the absence of the back-translation loss, a remarkable fall in the CIDEr score is observed in this method. In some cases, the method produces captions that are not related to the images. The method also produces the same caption for different images in some cases. The result analysis of this method proves that unpaired stylized text must be pre-trained in order to get good parametric performances.

3.2.5 Transformer and Graph-Based Methodology

There are many preexisting methods related to Transformer and Graph-Based Methodology [43], but a generic architecture of this model is considered here to demonstrate its working principle as described in the Algorithm. 18 and Figure. 20.

Algorithm 18 Transformer and Graph-Based Methodology

Require: Query Image (I)

Ensure: Word (Text representation, T)

- 1: Consider the query image I
 - 2: Consider the graph-structured scene descriptor as layer 1
 - 3: Send the output layer 1 to the SGtransformer layer ($SGTL$)
 - 4: $SGTL$'s output is transmitted to the Layout Analysis Layer (LAL)
 - 5: LAL 's output is represented by a vector (v)
 - 6: Transmit v to the image transformer layer (ITL)
 - 7: Perform auto-regressive sampling using ITL
 - 8: Train $VQVAE$ decoder
 - 9: Generate T using $VQVAE$ decoder
-

The methodology's block diagram is elucidated in Figure. 20 as follows:

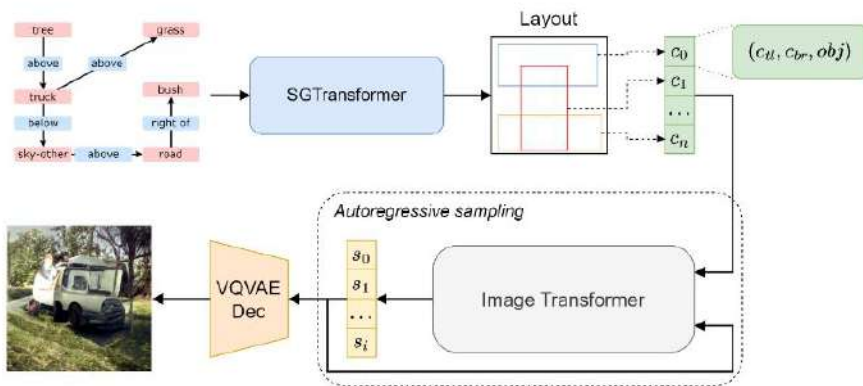


Fig. 20: Shows a Transformer and Graph-Based Methodology as proposed by [44]

Transformer and graph-based methodologies have multiple limitations. Some of the transform and graph-based methodologies have been applied to a limited number of datasets. Hence, they have failed to produce their versatility as far as application to multiple datasets is concerned. If the GNN module is removed from the proposed transformer and the graph-based method [45]

then the performance of the proposed method deteriorates drastically. The effectiveness of these methodologies can be enhanced through the application of the fusion method. Some of the methods are taking a huge amount of time and are burdensome. In some of these methods, more inductive bias can be introduced to achieve a more economical model.

3.2.6 Graph-Based Methodology

There are many preexisting methods related to Graph-Based Methodology [46], but a generic architecture of this model is considered here to demonstrate its working principle as described in the Algorithm. 19 and Figure. 21.

Algorithm 19 Graph-Based Methodology

Require: Query Image; I

Ensure: Word (Text representation)

- 1: Consider the query image I
 - 2: Consider the graph-structured scene descriptor layer 1
 - 3: Send the output layer 1 to the SGtransformer layer
 - 4: Output of the SGtransformer layer sent to layout analysis layer
 - 5: Output of layout analysis layer which is considered to be a vector sent to the image transformer layer and used for auto-regressive sampling
 - 6: Train VQVAE decoder to generate the output caption
-

The methodology's block diagram is elucidated in Figure. 21 as follows:

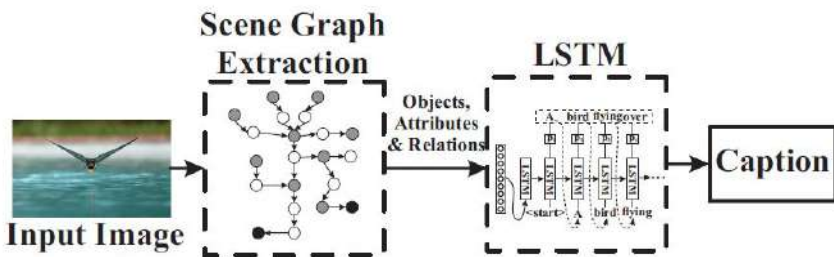


Fig. 21: Shows a Graph-Based Methodology as proposed by [46]

The performance and accuracy of these models can be improved, by adjusting the weights of the edges. These types of models are not applied in mixed-media databases. These types of models are not able to produce cross-modal correlations.

3.2.7 Attention and Graph-Based Methodology

There are many preexisting methods related to Attention and Graph-Based Methodology [47], but a generic architecture of this model is considered here to demonstrate its working principle as described in the Algorithm. 20 and Figure. 22.

Algorithm 20 Attention and Graph-Based Methodology

Require: Query Image (I)

Ensure: Word (Text representation, T)

- 1: Consider I as an input to the model
 - 2: Determine the scene graphs (sg) from I
 - 3: Construct sub-graph (ssg) set from sg
 - 4: Capture a semantic component (sc) from each sg
 - 5: Construct a sub-graph proposed network ($SGPN$)
 - 6: Identify usable/selected sg using $SGPN$
 - 7: Create an LSTM+Attention Based Model ($LABM$)
 - 8: Decode each selected sg using $LABM$
 - 9: Generate a sentence using $LABM$
-

The methodology's block diagram is elucidated in Figure. 22 as follows:

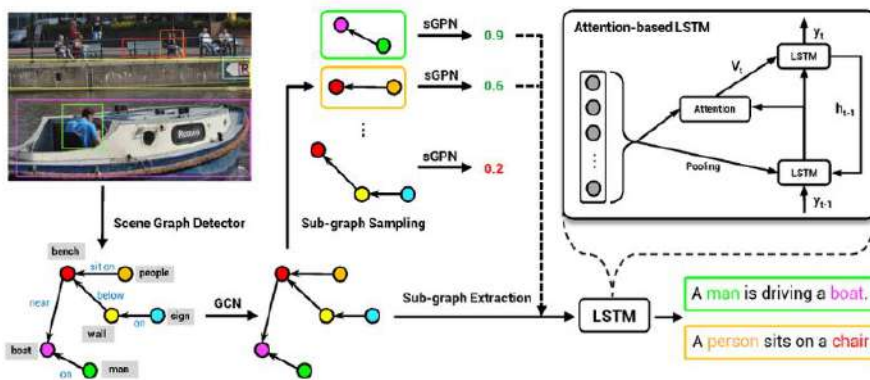


Fig. 22: Shows an Attention and Graph-Based Methodology as proposed by [47]

The Attention and Graph-Based methods have multiple limitations. These methods face lots of challenges related to the scene graph. These methods face lots of difficulties related to the extraction of the scene graph. Certain methods lack the ability to establish relationships between objects within the query image, and their graph parsers may remain unimproved.

3.2.8 CNN-Based Methodology

There are many preexisting methods related to CNN-based methodology [48], but a generic architecture of this model is considered here to demonstrate its working principle as described in the Algorithm. 21 and the Figure. 23.

Algorithm 21 CNN-Based Methodology

Require: Query Image (I)

Ensure: Word (Text representation, T)

- 1: Consider I as an input to the framework
 - 2: Construct the framework using 4 modules as follows: (a) vision (vm), (b) language (lm), (c) attention (am), and (d) prediction (pm)
 - 3: Use VGG-16 layer to construct vm
 - 4: Use RNN+CNN without max-pooling layer to construct lm
 - Perform context memorization using RNN
 - Perform context modeling using CNN
 - 5: Use Hidden Neural Network Layer ($HNNL$) to construct pm
 - 6: Calculate each lm level's attention vectors (av) using am
 - 7: Predict T using pm
-

The methodology's block diagram is elucidated in Figure. 23 as follows:

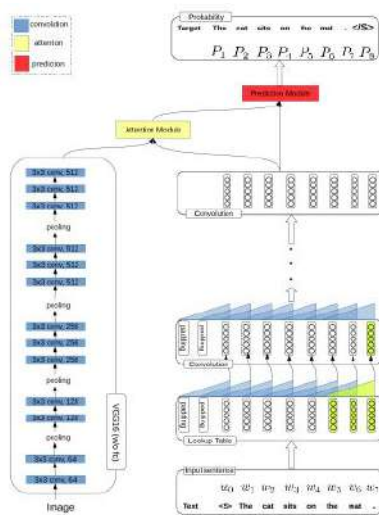


Fig. 23: Shows a CNN-Based Methodology as proposed by [48]

These types of methods suffer from many limitations. Many of these methods are not applied and tested to multiple datasets. Many of these methods are not tested based on various performance. In this type of method, it has

been observed that if the kernel width or the network depth improves then the complexity of these methods increases. These type of methods suffers from heavy training time and data overfitting problem.

3.2.9 Unsupervised and Reinforcement Learning-Based Methodology

There are many preexisting methods related to Unsupervised and Reinforcement Learning-Based Methodology [27], but a generic architecture of this model is considered here to demonstrate its working principle as described in the Algorithm. 22 and the Figure. 24.

Algorithm 22 Unsupervised and Reinforcement Learning-Based Methodology

Require: Query Image (I)

Ensure: Word (Text representation, T)

- 1: Consider I as an input to the framework
 - 2: Construct an Image Encoder Model ($ImgEM$) using (Convolutional + ReLU) + Max-Pooling + (Fully Connected + ReLU) + Softmax Layer
 - 3: Insert I into $ImgEM$
 - 4: Construct a Discriminator Model ($DisM$) which act as follows:
 - Consider T as an input to $DisM$
 - Discriminate real and fake captions
 - Generate a reward based on the real and fake captions
 - 5: Construct a Caption Generator Model ($CapGM$)
 - Consider $ImgEM$'s output as an input to $CapGM$
 - Produce T from the reward of $DisM$
-

The methodology's block diagram is elucidated in Figure. 24 as follows:

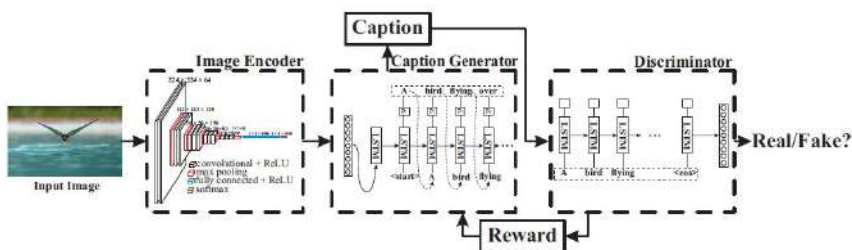


Fig. 24: Shows a Unsupervised and Reinforcement Learning-Based Methodology as proposed by [27]

Due to several limitations of the supervised learning-based image captioning methods, such as difficulty, while preparing the datasets, heavy

training time, and critical training process, the unsupervised and reinforcement learning-based methods are preferred. These methods discourage reliance on specific methods and datasets for generating image-caption pairs, aiming to maintain the quality and quantity of such pairs.

3.2.10 Vision-Language Pre-Training (VLP)

There are many preexisting methods related to Vision-language Pre-Training (VLP) [49], but a generic architecture of this model is considered here to demonstrate its working principle as described in the Algorithm. 23 and the Figure. 25.

Algorithm 23 Vision-language Pre-Training (VLP)

Require: Query Image; I

Ensure: Word (Text representation)

- 1: Consider the query image I
 - 2: Insert the input query image into the CNN structure
 - 3: Consider a pre-generated caption to pre-train the model. Pass the pre-generated caption for text embedding
 - 4: Generate the output after text embedding and CNN structure. Insert these outputs into a transformer encoder structure generated earlier
 - 5: Feed transformer encoder's output is fed to *MaskLM* using the image text matching, and the transformer decoder constructed earlier
 - 6: Transformer decoder helps to perform the object detection and produces the final caption using caption generator
-

The methodology's block diagram is elucidated in Figure. 25 as follows:

These methods have various limitations. These methods have failed to create the fusion relationship between image and text from the query image. These methods have failed to establish image and language pre-training jobs. Some of these methods have failed to show accurate and high performance in terms of performance parameters.

3.2.11 Transformer Based Methodology

There are many preexisting methods related to transformer-based methodology [50], but a generic architecture of this model is considered here to demonstrate its working principle as described in the Algorithm. 24 and the Figure. 26.

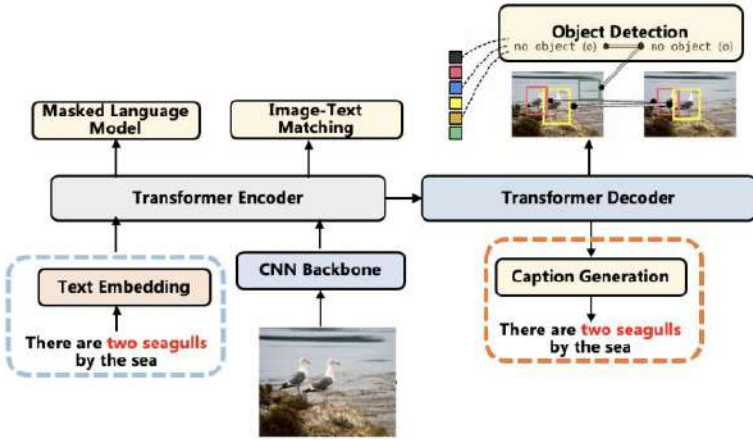


Fig. 25: Shows a Vision-language Pre-Training (VLP) as proposed by [49]

The equation used in algorithm 24 is explained as follows:

ky and vl vectors of MAL are constructed using Equation. 19 as follows:

$$ky_{t1} = W1_{ky} \cdot x1_{t1}, vl_{t1} = W1_{vl} \cdot x1_{t1}, \text{ and } qt_{t1} = W1_{q1} \cdot x\hat{1}_{t1} \quad (14)$$

where, $W1_{ky}$, $W1_{vl}$ and $W1_{qt}$ are transformation matrices

The final result $y1_{t1}$ can be generated using Equation. 20 as follows:

$$y1_{t1} = \sum_{ct=1}^{N1} a1_{tc1} \cdot vt_{ct} \text{ where, } a1_{tc1} = \frac{\exp(\frac{q1_{t1}^{T1} \cdot vt_{ct}}{\sqrt{d_{h1}}})}{\sum_{i1} \exp(\frac{q1_{t1}^{T1} \cdot vt_{ct}}{\sqrt{d_{h1}}})} \quad (15)$$

where, $(\frac{q1_{t1}^{T1} \cdot vt_{ct}}{\sqrt{d_{h1}}})$ is the a scaled dot-product between the two vectors, vt_{ct} is the average vector values of each head, d_{ht} is the dimension of each head which is calculated as $d_{h1} = \frac{d}{H1}$ where, $H1$ is the total number of heads

$PWFFL$'s output is computed using Equation. 21 as follows:

$$FF1(x1_{t1}) = \cup_{\sigma}(V1 \cdot x1_{t1} + b1) + ct \quad (16)$$

where, $\sigma(x1) = \max(x1, 0)$ is the ReLU activation function, $V1$, and $U1$ are learnable weight matrices, and $b1, ct$ are bias terms.

The add-norm operation performed by the Skip connection and normalization layer is defined by,

$$AddNorm1(x1_{t1}) = LayerNorm1(x1_{t1} + f1(x1_{t1})) \quad (17)$$

Algorithm 24 Transformer Based Methodology

Require: Query Image (I), Caption Path($CapP$), Total-Image-Number(n)

Ensure: Word (T)

- 1: Consider I , and $CapP$
 - 2: Count n from the dataset (ds)
 - 3: Construct an Image Feature Extraction Model ($ImgFEM$) using a deep neural network such as inception V3
 - 4: Use Software Layer (sl) to avoid image classification
 - 5: Using the preprocessing step convert all images into the same size.
 - 6: Feed all images into the framework
 - 7: Perform positional encoding
 - 8: Calculate the attention weights qt, ky , and vl
 - 9: Construct an Encoder-Decoder Layer (EDL)
 - 10: Construct a Transformer Layer (TFL) using the following layers:
 - Construct a Multi-Attention Layer (MAL) using linear transformation as described in Eq. 19 and 20
 - Construct a Position-Wise Feed-Forward Layer ($PWFFL$) using Eq. 21
 - Construct a Skip Connection + Normalization Layer ($SCNL$) using Eq. 22
 - 11: Define Hyper-Parameters (hp) to train TFM
 - 12: Train the Defined Model (DM) using hp
 - 13: Evaluate the performance value using any performance parameters such as BELU
-

where, $AddNorm1(xt1)$ is the normalization addition function over $x1_{t1}$, $LayerNorm1()$ is the normalization function of a given layer, and $f1$ is either an al or a $PWFFL$.

The methodology's block diagram is elucidated in Figure. 26 as follows:

Despite multiple advantages, such as solutions to the vanishing gradient and sequential execution problems, transformer-based methodologies face various challenges. These methods can not address the lack of context problem. These methods require huge computational costs while training the data. The performance of these methods is significantly impacted if the dataset includes duplicate query images. Dataset creators should ensure that captions do not rely on relative positional words for description, employing techniques like data augmentation to achieve this.

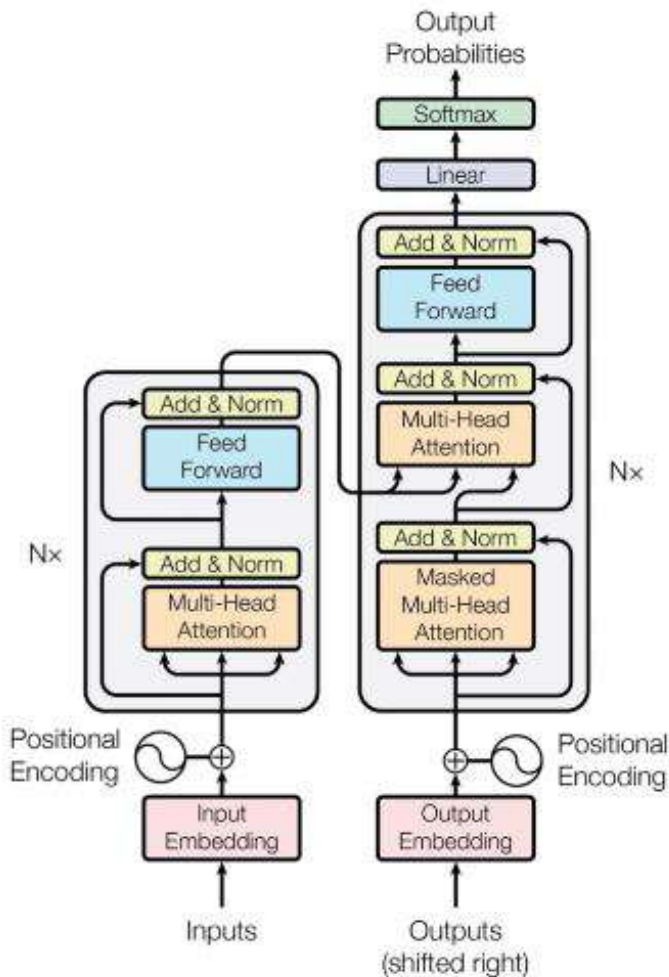


Fig. 26: Shows a Transformer Based Methodology as proposed by [50]

After describing, and analyzing various categories of image caption generation methodologies, it is important to execute different methods using various datasets and various performance parameters. This is described in the next section.

4 Result and Discussion

The performance of these methods is significantly impacted if the dataset includes duplicate query images. Dataset creators should ensure that captions do not rely on relative positional words for description, employing techniques

like data augmentation to achieve this. In this section, we conduct a comparative analysis of various pre-existing approaches, focusing on different KPIs outlined in section 2. Here, we briefly outline the hardware and software requirements for the experiments, as well as the multiple datasets used in the image captioning model. Additionally, we describe various performance parameters employed to gauge the efficacy of different methods. Towards the conclusion of this section, we present descriptions of multiple resulting tables and graphs, offering insights into various methodologies proposed over the past few decades.

4.1 Hardware and Software Requirements of the Experiments

Various image caption generation methodologies are simulated and tested for producing results using a desktop computer with the following hardware configuration. These elements include Intel Core i7-10th generation processor, 16GB DDR3 primary memory (RAM), a 512GB Kingston SATA-III SSD internal hard disk, a 5TB Western Digital external memory, an Intel HD internal 2GB Graphics Card, a 4GB NVIDIA external Graphics Card, a 17 inch Lenovo LCD Monitor, a digital Frontech Keyboard, and an optical Frontech Mouse. For implementing a huge amount of results for a prolonged period of time a constant power supply is provided.

The image caption generation methodologies are simulated and tested for producing results using a desktop computer with the following software configuration. The software configuration includes Windows 10 as the Operating System. To evaluate the proposed approach, we extensively tested it on various images from the MS-COCO dataset, employing Anaconda as a Python distributor. Using the Anaconda Navigator, we executed a base environment, enabling access to an open web application source to run program known as Jupiter Notebook. Origin-Pro 8.5 is used for generating the comparative graphs.

4.2 Datasets

Various image caption generation methods use various datasets for performance testing and analysis. Among many datasets, very few datasets are considered to be robust, versatile, relevant, effective, and efficient. Only very few datasets consist huge number of images. Only a few datasets have a variety of images. MS-COCO and Flickr datasets are considered to be versatile, huge, and effective. However, in order to understand the description various datasets used by various image caption generation methods are described in the section below.

4.2.1 MS-COCO

This dataset [51] was developed by Microsoft Corporation for the detection of a large-scale object, for the segmentation of objects, and to perform image

caption. The dataset contains a feature of in-depth instance annotations and object segmentation, along with features like super-pixel stuff segmentation. It comprises over 330,000 labeled images, encompassing 1.5 million object instances and featuring 5 captions for each image.

4.2.2 Flickr

The Flickr30k [52] dataset has firmly established itself as the gold standard for generating sentence-based image descriptions. It introduces Flickr30k Entities, an expansion of this dataset, which significantly enhances the existing pool of 158,000 captions from Flickr30k by incorporating 244,000 coreference chains. These chains establish connections between identical entities mentioned across different captions associated with the same image, all while being linked to 276,000 meticulously annotated bounding boxes. This annotation proves to be of paramount importance in advancing both automatic image description and our fundamental understanding of language. It paves the way for the creation of a novel benchmark, enabling precise localization of textual entity mentions within images.

4.2.3 The Remote Sensing Image Captioning Dataset (RSICD)

This dataset [53] can generate image captions for over 10,000 remote sensing images gathered from diverse sources such as Google Earth. Each image in this dataset is sized at 224x224 pixels and can generate descriptions comprising five sentences.

4.2.4 Oxford 102

This dataset [54] is assembled by a comprehensive dataset featuring 102 distinct flower categories, each commonly found in the United Kingdom. The dataset consists of different flower classes, with each class containing a varying number of images, spanning from 40 to 258. This encompasses images that exhibit substantial variations in scale, pose, and lighting. Furthermore, some categories display significant diversity within their own class, while others closely resemble each other. To facilitate visualization, isomap techniques are employed, considering both shape and color features in our dataset representation.

4.2.5 InstaPIC

This [55] dataset was compiled by gathering Instagram posts, totaling 721,176 pairs of images and captions contributed by 4.8k users. Employing a selection of 270 specific hashtags, the dataset creators utilized Instagram APIs to filter and retrieve posts, including both images and captions. However, a notable challenge with the InstaPIC dataset arises from the potential disparity between the captions and the actual image content. This disparity results from the

nature of user-generated captions on Instagram, which can often be vague and may not precisely describe the visual elements of the image.

4.2.6 YFCC

The YFCC dataset [56] is an extensive huge number of media assets, including millions of photographs and videos. All items in this dataset are accompanied by a Creative Commons license. Each media object is associated with a rich set of metadata, including details such as the Flickr identifier, owner's name, camera information, title, tags, geographical data, and media source.

4.2.7 Stanford Image Paragraph Captioning

This dataset [57] represents a subset extracted from the Visual Genome dataset, encompassing around 20,000 images that are meticulously paired with their corresponding textual paragraphs. It serves as a valuable resource specifically designed to support tasks related to image paragraph captioning. The dataset contains 19561 distinct data out of which, the training set true and false values are 14.6K and 4982 respectively, and the testing set true and false values are 2492 and 17.1K respectively.

4.2.8 IU X-Ray

The IU X-ray dataset, as cited in [58], comprises 3,955 images obtained from patient hospital databases of Indiana. Moreover, this dataset includes 7,470 chest X-rays obtained from hospital picture archiving systems. Each X-ray is paired with images showing both frontal and lateral views. The reports are structured with sections for comparison, indication, findings, and impression.

4.2.9 MIMIC-CXR

This extensive public dataset [59]. comprises chest radiology images (CT/MRI) stored in DICOM format. The images do not include any text related to radiological reports. The dataset includes approximately 377,110 captions, with 227,835 images gathered from a variety of radiographic studies.

4.2.10 IVKD

The Image Visual Keyword Dataset (IVKD), as cited in [14], originates from the MSCOCO dataset. It involves the categorization of 410 fine-grained categories, an expansion from the original 80 object categories found in MSCOCO. As a result, it creates an image vocabulary comprising 1,044 distinct categories. This dataset comprises five sets of images, each obtained through individual manual annotations utilizing a keyword extractor. Certain sets of images are omitted when no corresponding object labels are present. These exclusions are primarily due to limitations in object detection technology and the constraints of the keyword vocabulary. In total, IVKD encompasses 110,535 sets

of images for training purposes, as well as an additional 4,736 and 4,831 sets for validation and testing, respectively.

4.2.11 Chinese Image Caption (AIC-ICC)

The Chinese Image Caption dataset [18], also known as AIC-ICC, boasts an impressive collection of 300,000 images. Given its substantial scale, this dataset proves exceptionally versatile. This one is appropriate for both testing and analytical endeavors of a large scale. It was originally introduced by AI Challenger in 2017 and stands as the world's largest Chinese image caption dataset. Within this dataset, each image is accompanied by multiple Chinese descriptions, totaling a remarkable 1.5 million unique descriptions. Significantly, this dataset includes adjectives and Chinese idioms, enabling the modification of characters and scenes depicted in the images.

4.2.12 Labeled Faces in the Wild (LFW)

Labeled Faces in the Wild [60] serves as a public benchmark for face verification, often referred to as pair matching. An important point to be noted is that, the performance of an algorithm on LFW should not be solely relied upon to determine its suitability for commercial applications. This database comprises a collection of face photographs specifically curated for the exploration of unconstrained face recognition challenges. The dataset includes over 13,000 facial images sourced from the internet, each accompanied by the individual's name. Notably, 1,680 individuals within the dataset have two or more distinct photos represented.

4.2.13 News image dataset

This is a custom but private dataset [22] prepared from news broadcast videos. The articles are acquired from TIME magazine's website, making sure that each downloaded article includes one image along with an associated caption. The dataset comprises a total of 19,841 article-image-caption combinations. Additionally, the article titles and their corresponding keywords are collected. TIME magazine categorizes articles into ten different news categories, and the information about the specific category for each article is preserved. The vocabulary set for articles consists of 6,350 words, each with a frequency of over 100. In the case of captions, their vocabulary set includes 1,937 words, each occurring at least 20 times. When these two sets are merged, then a combined vocabulary of 6,449 unique words is obtained. Furthermore, the dataset encompasses 10 distinct news categories, along with a total of 719 unique keywords.

4.2.14 Sports video dataset

This dataset, as cited in [24], is a custom collection meticulously curated from sports broadcast videos, and it remains private. The dataset comprises 22

videos with a cumulative time span of 10 hours. Each of the videos has a frame resolution of 640 x 480 pixels and a frame rate of 25 fps. They are categorized into various genres and encompass content from multiple prominent broadcasters.

The dataset details are provided in Dataset Collection Table I, which explains the dataset name, domain name, no. of images, no. of videos, no. of objects, no. of classes, no. of captions.

Table 2: Dataset Collection Table I

Serial Number	Dataset Name	Domain Name	#images	#videos	#objects	#classes	#captions
1	MS-COCO	Open/Public	330000	-	1.5M	91	7.5 millions
2	Flickr	Open/Public	31,783	-	8.7	44,518	158000
3	RSICD	Open/Public	10921	-	240-1031	30	54605
4	Oxford 102	Custom	>4080	-	40-258	102	-
5	InstaPIC	Custom	1.1M	-	-	270	1 per pair (721,176 pairs)
6	YFCC	Open/Public	0.99M	0.8	100M	-	-
7	Stanford Image Paragaph Captioning	Open/Public	19561	-	35 per image	26 per image	-
8	IU X-Ray	Open/Public	3,955	-	-	-	-
9	MIMIC-CXR	Open/Public	227,835	-	6500	10	377,110
10	IVKD	Custom	110,535 sets	-	80	410	-
11	Chinese Image Caption	Public/Open	300,000	-	-	-	1.5 million
12	LFW	Public	13,000	-	-	-	-
13	News Image Dataset	Private	19,841	-	-	-	-
14	Sports Video Dataset	Custom	-	20	-	-	-

4.3 Performance Metrics

The accuracy and perfection in implementation is crucial to be evaluated for the image caption generation process. The majority of image captioning techniques utilize metrics such as ROUGE, BLEU, CIDEr, METEOR, and SPICE to assess their performance. But some of the methods [7] [24][22] [15] are used to recall, precision, accuracy score, and F1 score to calculate the performance. It has been observed that some of the methods [29] measure their performances using cross-entropy. Some of the existing methodologies [6] introduce Plausibility, and Relevance to measure the performances. Some established equations can be used to calculate these metrics. Sections 4.3.1 to 4.3.12 describe the analysis of various performance parameters associated with image caption generation.

4.3.1 BLEU

The BLEU [61] stands for Best Linear Unbiased Estimator. In this parameter, minimum variance is referred with the word best. Equation. 23 helps to compute the BLEU parameter.

$$\beta \leftarrow \begin{cases} 1 & \text{if } \mu > \theta \\ \varepsilon^{1-\frac{\theta}{\mu}} & \text{if } \mu \leq \theta \end{cases} \quad (18)$$

where,

β represents the brevity penalty, μ the candidate translation length, θ for effective reference corpus length and θ for residuals.

$$B \leftarrow \beta \cdot \exp^{\sum_{i=1}^k \omega_i \log_{10} \rho_i} \quad (19)$$

where,

β is brevity penalty, \exp the exponential, ρ_i the i -gram precisions, ω_i the i -th positive weight, i -th-gram, k -th-gram and the B represents BLEU Parameter.

4.3.2 METEOR

The METEOR [62] metric is used to overcome the drawbacks of the BLEU metric. The calculation can be done as the following:

$$Q_\alpha \leftarrow \frac{10 \cdot s \cdot n}{s + 9 \cdot n} \quad (20)$$

Where,

Q_α is functional mean, s denotes unigram recall, n for unigram precision and $s + 9n$ is the harmonic-mean.

If, pc is the number of possible chunks, un is the number of the unigrams matched and Pn Penalty;

Then, Pn can be calculated as,

$$Pn \leftarrow 0.5 * \frac{pc}{un} \quad (21)$$

With the help of the functional mean (Q_α) and penalty (Pn), we can calculate the METEOR Score (M_s) for the given alignment.

$$M_s \leftarrow Q_\alpha(1 - Pn) \quad (22)$$

4.3.3 ROUGE

This parameter [63] is used to calculate text summaries. This parameter can be calculated as ro-i, ro-l, and ro-s.

r-i is a parameter that is denoted by i-th-gram recall between the reference and the candidate. ro-i is calculated as follows:

$$ro - i \leftarrow \frac{\sum_{\dot{A} \in ref\ summaries} \sum_{u_i \in \dot{A}} C_m(u_i)}{\sum_{\dot{A} \in ref\ summaries} \sum_{u_i \in \dot{A}} C(u_i)} \quad (23)$$

where,

$ro - n$ is the parameter of Rouge for n -th-gram recall, i is the length of the i -th-gram, u_i for i -th-gram, $C_m(u_i)$ is the *maximum number of i -th - grams in the candidate summaries* and \dot{A} is the ref summaries set.

ro-l parameter is denoted by LCS-based data and computed as follows:

$$ro - l(\mathbb{R}, \lambda) \leftarrow (1 + \delta^2) \cdot r^{lcs}(\mathbb{R}, \lambda) \cdot p^{lcs}(\mathbb{R}, \lambda) \cdot r^{lcs}(\mathbb{R}, \lambda) + \delta^2 \cdot p^{lcs}(\mathbb{R}, \lambda) \quad (24)$$

where,

δ is recall or precision aspect, $ro - l(\mathbb{R}, \lambda)$ is the metric between a candidate document and a reference document, \mathbb{R} is the candidate document, λ is single reference document, $r^{lcs}(\mathbb{R}, \lambda)$ is the LCS set r-score, and $p^{lcs}(\mathbb{R}, \lambda)$ is the LCS set p-score.

ro-s parameter is denoted by i -gram with skips and computed as follows:

$$ro - s(\beta, \lambda) \leftarrow \frac{((1 + \delta^2) \cdot r_s(\beta, \lambda) \cdot p_s(\beta, \lambda))}{(r_s(\beta, \lambda) + (\delta^2) \cdot p_s(\beta, \lambda))} \quad (25)$$

where,

δ is the relative importance of the PR and RR, $ro - s(\beta, \lambda)$ is an F-score measure between a candidate document and a single reference document, β is candidate document, λ is the single reference document, $r_s(\beta, \lambda)$ is the skip-bigram set recall score, and $p_s(\beta, \lambda) \leftarrow$ the skip-bigram set precision score.

4.3.4 CIDEr

This parameter [64] can be calculated using the following formula:

$$C - par(v_m, B_m) \leftarrow \frac{1}{k} \sum_n \left(\frac{f^l(v_m) \cdot f^l(B_{mn})}{\|f^l(v_m)\| \cdot \|f^l(B_{mn})\|} \right) \quad (26)$$

Where,

$C - par(v_m, B_m)$ is the score for l -grams of a given length l is established by computing the average cosine similarity between the candidate and the reference sentences. $f^l(v_m), f^l(B_{mn})$ are vectors and $\|f^l(v_m)\|, \|f^l(B_{mn})\|$ is the magnitude of the vectors

4.3.5 SPICE

This parameter [40] can be calculated using the following formulas:

$$p(\alpha, \beta) \leftarrow \frac{tmp(\alpha) \cdot \theta tmp(\beta)}{tmp(\alpha)} \quad (27)$$

$$r(\alpha, p(\alpha, \beta) \leftarrow \frac{tmp(\alpha) \cdot \theta tmp(\beta)}{tmp(\alpha)}) \leftarrow \frac{tmp(\alpha) \cdot \theta tmp(p(\alpha, \beta) \leftarrow \frac{tmp(\alpha) \cdot \theta tmp(\beta)}{tmp(\alpha)})}{tmp(p(\alpha, \beta) \leftarrow \frac{tmp(\alpha) \cdot \theta tmp(\beta)}{tmp(\alpha)})} \quad (28)$$

$$s - param(\alpha, p(\alpha, \beta) \leftarrow \frac{tmp(\alpha) \cdot \theta tmp(\beta)}{tmp(\alpha)}) \leftarrow \frac{(2 \cdot p(\alpha, \beta) \cdot r(\alpha, \beta))}{(p(\alpha, \beta) \cdot r(\alpha, \beta))} \quad (29)$$

where,

α is the caption for a candidate, β is reference caption set, δ is the caption to the tuple mapping function, $s - param(c, R)$ is the spice parameter with argument α and β .

4.3.6 Precision Rate

It can be calculated [65] as follows:

$$precision - rate \leftarrow \frac{t_p}{(t_p + f_p)} \quad (30)$$

Where,

$t_p \leftarrow$ The count of samples correctly identified as positives.

$f_p \leftarrow$ The count of samples correctly identified as negatives.

4.3.7 Recall Rate

It can be calculated [65] as follows:

$$recall - rate \leftarrow \frac{t_p}{t_p + f_n} \quad (31)$$

Where,

$t_p \leftarrow$ The count of samples correctly identified as positives.

$f_n \leftarrow$ The no. of samples classified as negatives when they are actually positives.

4.3.8 F1-Score

It can be calculated [65] as follows:

$$F1 - Score \leftarrow \frac{2 \cdot (p \cdot r)}{p + r} \quad (32)$$

Where,

$p \leftarrow$ the precision rate

$r \leftarrow$ the recall rate

4.3.9 Accuracy Score

It can be calculated [65] as follows:

$$accuracy - score \leftarrow \frac{(total\ no\ of\ correct\ prediction)}{(total\ no\ of\ prediction)} \quad (33)$$

4.3.10 Cross-Entropy Loss

It [66] can be computed in a subsequent manner:

$$h(\theta, \phi) \leftarrow \sum_{i \in I} \theta(i) \cdot \log(\phi(i)) \quad (34)$$

where,

$h(\theta, \phi) \leftarrow$ The value associated with the probabilities of events derived from θ and ϕ

$\theta(i) \leftarrow$ The likelihood of event i occurring within the context of θ

$\phi(i) \leftarrow$ The likelihood of event i occurring within the scope of ϕ

\log the base-2 logarithm

4.3.11 Plausibility

It [67] can be computed in a subsequent manner:

$$pl(\tau) \leftarrow 1 - \omega(\bar{\tau}) \quad (35)$$

Where,

$pl() \leftarrow$ the plausibility function

$\omega() \leftarrow$ the belief function

$\tau \leftarrow$ A portion of an enumerable set X

$\bar{\tau} \leftarrow$ complement of τ

4.3.12 Relevance Score

The relevance score [68] is very important to calculate image caption generation. In any case it is determined by a system's ability to search for a keyword within a collection of written texts and assign relative scores to the results that successfully match the searched keywords. This is established using a set of specific standards, and a few of these specific standards include:

- (a) No instances of the keyword match in whole text.
- (b) Is the title is present in the keyword?
- (c) Is the keyword present in the abstract?
- (d) Is the keyword present in the title and/or abstract?

The relevance score represents a value. If it is high, it means that the checked keyword is more important for the results. This score can be calculated using Eq. 41 as follows:

$$Rel - Score \leftarrow \frac{1}{m} \cdot \sum_{i=1}^m cosine - similarity(x_a^b \cdot y_a) \quad (36)$$

where, $x_a^b \leftarrow$ denotes context features of the candidate sentence

$y_a \leftarrow$ denotes image features

Performance Details Table I shows different parameter names, descriptions, variations, equations, and remarks associated with the performance analysis of the different image caption analysis methodologies.

Table 3: Parameter Details Table I

Parameter Name	Description	Variations	Equation	Remark
BLEU	Bilingual Evaluation Understudy	BLEU-1, BLEU-2, BLEU-3, BLEU-4	$B \leftarrow \beta \cdot \exp \sum_{i=1}^k \omega_i \log_{10} \rho_i$	machine-generated texts
METEOR	Metric for Evaluation of Translation with Explicit Ordering	-	$M_{Score} \leftarrow f_{\alpha}(1 - P)$	compares word segments
ROUGE	Recall-Oriented Understudy for Gisting Evaluation	ROUGE-N, ROUGE-L, ROUGE-W, ROUGE-S, ROUGE-SU	$r - l(\alpha, \beta) \leftarrow (1 + \delta^2) \cdot r^{lcs}(\alpha, \beta) \cdot p^{lcs}(\alpha, \beta) \cdot r^{lcs}(\alpha, \beta) + \delta^2 \cdot p^{lcs}(\alpha, \beta)$	quality of text summarization
CIDEr	Consensus-based Image Description Evaluation	CIDEr-D	$c - \text{param}(c_m, S_m) \leftarrow \frac{1}{k} \sum_n \left(\frac{f^l(c_m) \cdot f^l(s_{mn})}{\ f^l(c_m)\ \cdot \ f^l(s_{mn})\ } \right) - \text{param}(\alpha, p(\alpha, \beta)) \frac{tmp(\alpha) \cdot tmp(\beta)}{tmp(\alpha)}$	explicit evaluation for image captions based on semantic context
SPICE	Semantic Propositional Image Caption Evaluation	-	$\text{precision} - \text{rate} \leftarrow \frac{t_p}{(tp+fp)}$	count of samples correctly identified as positives and negatives
Precision Rate	Precision Rate for Image Caption Evaluation	-	$\text{recall} - \text{rate} \leftarrow \frac{t_p}{tp+fn}$	count of samples correctly identified as positives and false negatives
Recall Rate	Recall Rate for Image Caption Evaluation	-	$F1 - \text{Score} \leftarrow \frac{2 \cdot (p \cdot r)}{p+r}$	based on precision rate and recall rate
F1-Score	F1-Score for Image Caption Evaluation	-	$\text{accuracy} - \text{score} \leftarrow \frac{\text{score}}{(\text{total no of correct prediction})}$	based on total no. of correct prediction and total no. of prediction
Accuracy Score	Accuracy Score for Image Caption Evaluation	-	$h(\theta, \phi) \leftarrow \sum_{i \in I} \theta(i) \cdot \log(\phi(i))$	based on likelihood of event i occurring within the context
Cross-Entropy Loss	Cross-Entropy Loss for Image Caption Evaluation	-	$p(\tau) \leftarrow 1 - \omega(\bar{\tau})$	based on belief function, an enumerable set
Plausibility	Plausibility for Image Caption Evaluation	-	$\text{Rel} - \text{Score} \leftarrow \frac{1}{m} \cdot \sum_{i=1}^m \text{cosine} - \text{similarity}(x_a^b \cdot y_a)$	keyword searching on the entire written text
Relevance Score	Relevance Score for Image Caption Evaluation	-		

4.4 Performance Analysis

This section represents Resultant Tables I and II which showcases the performances of all the existing methodologies as discussed in Comparative Table - I based on BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE, CIDEr, and SPICE parameters. This section represents Resultant Tables III and IV which showcase the performances of all the existing methodologies as discussed in Comparative Table - I based on PR, RR, FS, and AS parameters. To evaluate the performance of the Resultant Tables MS-COCO and Flicker datasets are used. Resultant Tables I and III are implemented based on the MS-COCO datasets, whereas Resultant Tables II and IV are implemented based on the Flicker datasets. To showcase the pictorial representation of all Resultant Tables, Fig. 27 and 28 are represented. Fig. 27 ((a),(c),(e),(g)) and Fig. 28 (a) represent BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE, CIDEr, and SPICE parameter values based on the MSCOCO dataset. Fig. 28 (c) represents the PR, RR, FS, and AS parameter values based on the MSCOCO dataset. Fig. 27 ((b),(d),(f),(h)) and Fig. 28 (b) represent BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE, CIDEr, and SPICE parameter values based on the Flicker dataset. Fig. 28 (d) represents the PR, RR, FS, and AS parameter values based on the Flicker dataset. This section represents the Resultant Table V which shows the sample image, generated caption, and applied methodology during image caption generation.

Table 4: Resultant Table I

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr	SPICE
ALT	76.15	58.00	44.80	33.60	25.70	56.20	112.15	21.11
CSMN	59.10	62.25	43.19	22.73	27.14	58.90	123.13	-
TAAM	78.60	65.15	50.17	36.18	26.90	56.16	117.12	-
VSA	63.15	44.00	31.91	24.58	18.95	-	67.16	-
MULTIMODAL	49.93	32.19	22.91	18.17	21.16	42.73	117.85	-
CDEM	-	-	-	-	24.90	-	-	-
TOM	76.00	57.17	44.14	32.40	26.15	51.60	97.73	-
MLPN	81.20	64.13	47.91	35.70	27.30	58.10	113.80	-
CAVP	81.10	64.60	51.00	38.10	28.10	58.11	120.16	-
MLADIC	79.60	63.15	48.25	36.10	28.10	57.60	120.50	-
STACK-VS	78.90	63.45	48.90	37.30	27.80	57.50	118.90	-
MADASAP	80.75	65.13	51.10	38.65	29.17	58.51	129.18	21.80
VAM	75.20	56.19	43.44	32.75	25.59	-	96.00	-
STMA	80.13	64.75	49.85	37.70	28.30	58.10	122.95	-
GLDO	-	-	-	36.33	28.12	27.80	121.25	22.95
NADGCN	77.10	63.90	48.95	36.15	27.85	58.10	115.90	21.33
NICVATP2L	67.15	46.75	33.95	22.95	30.35	52.15	67.10	-
ARRGF	-	-	-	14.85	25.45	33.15	40.25	-
BDRGRUN	63.20	44.35	30.25	21.10	20.35	46.40	65.75	13.15
SCS	66.90	48.10	33.45	22.25	20.95	47.35	72.35	12.75

Table 5: Resultant Table II

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr	SPICE
ALT	77.15	58.10	46.25	34.10	27.15	55.20	111.15	20.20
CSMN	55.10	60.60	41.75	22.10	27.10	58.45	121.10	22.10
TAAM	80.40	65.30	49.15	37.65	27.15	57.60	116.55	20.60
VSA	61.60	45.75	32.65	23.75	20.75	56.75	66.80	22.80
MULTIMODAL	50.55	31.60	22.70	18.55	20.60	43.40	118.45	20.35
CDEM	-	-	-	-	25.65	55.60	105.30	20.75
TOM	74.40	56.30	43.35	31.55	25.10	53.10	98.60	21.75
MLPN	79.15	62.70	47.30	35.30	27.75	56.70	113.40	20.35
CAVP	80.75	63.60	50.70	37.40	28.75	58.45	120.55	20.60
MLADIC	78.30	63.35	45.25	36.55	28.70	56.65	120.40	20.45
STACK-VS	79.55	63.60	48.75	37.30	27.30	57.45	117.55	-
MADASAP	80.40	64.75	49.65	38.40	29.30	59.40	128.70	22.35
VAM	73.55	55.70	42.60	31.60	25.70	-	95.75	-
STMA	79.30	63.10	48.55	36.65	27.70	58.75	123.55	20.40
GLDO	-	-	-	36.40	27.55	27.70	121.75	20.30
NADGCN	77.75	64.65	48.40	35.30	27.60	57.40	115.85	21.40
NICVATP2L	67.55	46.45	34.45	23.60	30.75	52.40	66.40	20.85
ARRGF	-	-	-	14.55	25.80	32.55	40.55	20.60
BDRGRUN	63.55	44.45	30.45	20.60	20.70	46.55	65.40	12.60
SCS	66.60	46.40	32.50	22.85	21.65	47.40	72.75	12.30

From the Resultant Table I, Fig. 27 ((a),(c),(e),(g)) and Fig. 28 (a) it is clear that MADASAP [9], gives the highest performance in terms of BELU-1, BELU-2, BELU-3, BELU-4, ROUGE, CIDEr, and SPICE parameters. From the Resultant Table I, Fig. 27 ((a),(c),(e),(g)) and Fig. 28 (a) it is clear that NICVATP2L [18], gives the highest performance in terms of METEOR parameter. From the Resultant Table I, Fig. 27 ((a),(c),(e),(g)) and Fig. 28 (a) it is clear that the MULTIMODAL method generates the lowest value in terms of BELU-1, BELU-2, and BELU-3 parameters, the ARRGF generates the lowest value in terms of BELU-4, ROUGE, CIDEr parameters, the VSA generates the lowest value in terms of METEOR parameter, and BDRGRUN and SCS generate SPICE parameter.

From the Resultant Table II, Fig. 27 ((b),(d),(f),(h)) and Fig. 28 (b) it is clear that MADASAP [9], gives the highest performance in terms of BELU-1,

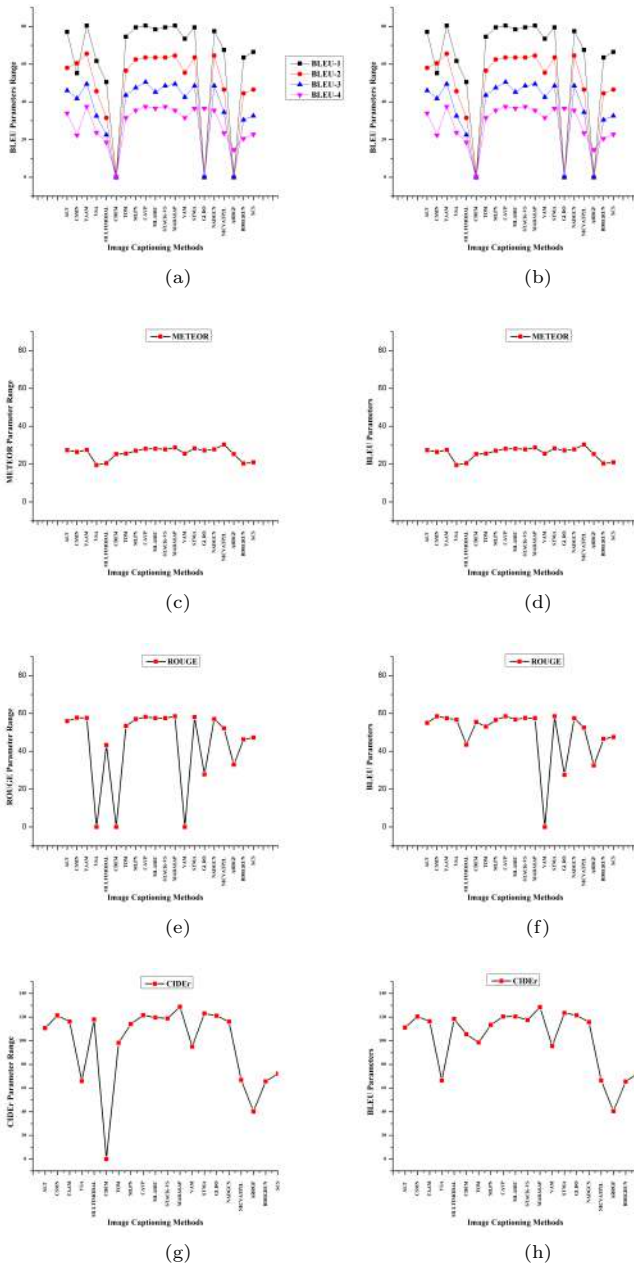


Fig. 27: Evaluation of Performance metrics based on MS-COCO and Flickr Datasets: (a) BLEU Matrix for MS-COCO dataset (b) BLEU Matrix for Flickr dataset (c) METEOR Matrix for MS-COCO dataset (d) METEOR Matrix for Flickr dataset (e) ROUGE Matrix for MS-COCO dataset (f) ROUGE Matrix for Flickr dataset (g) CIDEr Matrix for MS-COCO dataset (h) CIDEr Matrix for Flickr dataset

Table 6: Resultant Table III

Method	PR	RR	FS	AS
UCMA	76.52	77.31	71.93	76.90
GTSCD	97.45	95.65	-	96.65
CDEM	49.10	20.00	39.45	-
MADASAP	-	-	44.75	-
VSAM	91.75	89.03	-	76.00

Table 7: Resultant Table IV

Method	PR	RR	FS	AS
UCMA	75.55	75.55	71.60	76.75
GTSCD	98.60	95.40	71.10	96.45
CDEM	49.45	19.45	39.91	-
MADASAP	-	-	44.40	-
VSAM	90.90	89.55	45.60	75.40

BELU-2, BELU-3, BELU-4, METEOR, ROUGE, CIDEr, and SPICE parameters. From the Resultant Table II, Fig. 27 ((b),(d),(f),(h)) and Fig. 28 (b) it is clear that the MULTIMODAL method generates the lowest value in terms of BELU-1, BELU-2, and BELU-3 parameters, the ARRGF generates the lowest value in terms of BELU-4 parameter, the MULTIMODAL, and BDRGRUN generates the lowest value in terms of METEOR parameter, the GLDO method generates the lowest value in terms of ROUGE parameter, the ARRGF generates the lowest value in terms of CIDEr parameter, and BDRGRUN and SCS generate SPICE parameter.

From the Resultant Table III, IV, and Fig. 28 ((c),(d)) it is clear that GT + SCD Method [24] gives the best Precision Rate, Recall Rate, and Accuracy Rate whereas UCMA [7] gives the best F1 Score in comparison with other methods. From the Resultant Table III, IV, and Fig. 28 ((c),(d)) it is clear that CDEM generates the lowest value in terms of PR, RR, FS parameters, whereas VSAM generates the lowest value in terms of AS.

All results are computed with respect to the MS-COCO and Flickr standard datasets, and the dataset undergoes normalization, balancing, and standardization before being used. Table 5 showcases various methods from Resultant Tables I and II, displaying the resultant captions and the corresponding method names.

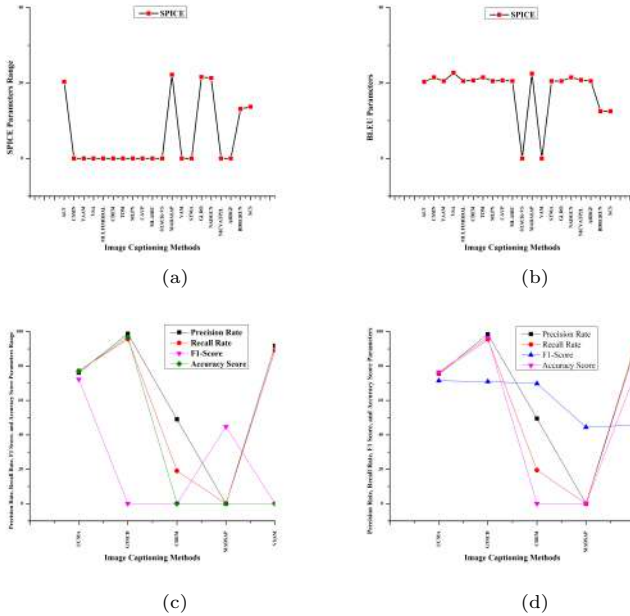


Fig. 28: Evaluation of Performance metrics based on MS-COCO and Flickr Datasets: **(a)** SPICE Matrix for MS-COCO dataset **(b)** SPICE Matrix for Flickr dataset **(c)** PR, RR, F1 Score, and AS Matrices for MS-COCO dataset **(d)** PR, RR, F1 Score, and AS Matrices for Flickr dataset

5 Limitations of the existing approaches

During image caption generation researchers and scientists face multiple problems such as explore bias (EBP), loss-evaluation mismatch (LEMP), vanishing gradient (VGP), exploding gradient (EGP), etc. Along with these problems, problems like OH, IC, CU, RE, etc. can also be encountered. In this section, some of the major image caption generation problems are briefly discussed, which is followed by brief discussions.

























5.1 Explore Bias (EBP)

This issue arises when the model trained solely on the training data lacks exposure to its own predictions. This problem often occurs in RNN models that predict the next word based on the previous one. Reinforcement learning-based optimization models can be used to address this issue.

5.2 Loss-Evaluation Mismatch (LEMP)

Sometimes during the testing of an image caption generation/language model, some evaluation matrices are used to measure the result. If these matrices are non-differentiable then they can not be used for result evaluation. Under

Table 8: Shows the generated captions from Resultant Table V

Sample Image				
Generated Caption	a bear laying on its back in the grass	beautiful style fashion stool	A man is standing in a store	boy is doing backflip on wakeboard
Applied Methodology	ALP	CSMN	TAAM	YSA
Sample Image				
Generated Caption	A bridge on a river with some green trees on two sides of it	BestBuy CEO Brian Dunn resigned amid investigation into his personal conduct	a young boy in a baseball uniform standing on a field	a yellow fire hydrant sitting on a sidewalk next to a fence
Applied Methodology	MULTIMODAL	CDEM	TOM	MLPN
Sample Image				
Generated Caption	a man is surfing in the ocean	two girls are playing a baseball game	a man wearing a suit and tie taking a picture of himself in a mirror	a man is playing basket ball
Applied Methodology	CAVP	MLAFCDIC	STACK-VS	MADASAP
Sample Image				
Generated Caption	a yellow and purple train is passing by	a lady is cleaning her teeth	a zebra standing in the snow next to a stone wall	a red motorcycle parked in a dirt field
Applied Methodology	VAM	STMA	GLDO	NADGCN
Sample Image				
Generated Caption	There is a man in glasses repairing the car in the room	Lung Lesion	a man on a court with a tennis racket	a cat sitting on top of a wooden bench
Applied Methodology	NICVATP2L	RGF	BDRGRUN	SCS
Sample Image				
Generated Caption	vice president Dick Cheney speaks at the luncheon ceremony	a bowler is taking his run-up	two children are holding tennis rackets and balls	a bunch of donuts on the table
Applied Methodology	UCMA	GTSCD	MADSAP	VSAM

these circumstances, LEM errors can be generated. This type of error can be overcome using reinforcement learning-based optimization models.

5.3 Vanishing Gradient (VGP)

This type of problem occurs in those ANNs where the network is trained using gradient or back-propagation. This error occurs when the weight updates of the neural network are determined by calculating the partial derivative of the error function with respect to the current value stored in each weight at

each step. Due to this type of problem, the image caption generation model's training operation may be permanently stalled.

5.4 Exploding Gradient (EGP)

This problem is associated with the gradient. In the context of deep or recurrent neural networks, there is a tendency for error gradients to accumulate and grow significantly during the update process. This can lead to the gradients reaching extremely high values. Consequently, the weights of the network experience substantial updates, causing the network to become highly unstable. In the most favorable scenario, a deep multi-layer Perceptron network may fail to learn from the training data, ultimately producing weight values that turn into NaN (Not a Number), rendering further updates impossible. Similarly, in recurrent networks, the phenomenon of exploding gradients can lead to an unstable network incapable of learning from the training data.

5.5 Object Hallucination (OHP)

In this problem, different absent objects in the input images are found during the image captioning process. The result of this problem leads to low quality image captioning results, which is not desired for visually challenged people. To measure this problem, the CHAIR metric is invented. This type of problem is solved using GAN-based models.

5.6 Illumination Conditions

If any image is captured during poor light or low illumination conditions or inside a house then this type of problem may occur. Due to this type of problem poorly constructed images are generated and poor and inaccurate captions are generated. Due to this type of problem image captioning model is not able to perform effectively. In order to overcome this type of problem various methodologies such as contrast enhancement, color correction, and low-light image enhancement may be incorporated.

5.7 Contextual Understanding

In case of this problem, the image captioning model struggles to grasp the context of the image, the relationship between the objects, and their spatial arrangement within the query image. Addressing this issue is particularly challenging.

5.8 Referring Expressions

In this problem, the image captioning method fails to establish the correct expression of an object present in the image. In this type of problem sometimes the image caption generation model either fails to point to the correct object from the query image or the model fails to establish the link between

the correct object with the correct expressions. This problem becomes harder to solve in case of multiple objects present in the query image having nearly same expression. To solve this problem the query image must be studied properly and the linguistic or visual characteristics of the query image must be understood properly.

6 Novelties and Advantages of the Current Article

The article demonstrates various methodologies related to image caption analysis (published in recent times), algorithms related to image caption generation, datasets, performance parameters, and performance analysis. After analyzing the above-mentioned ideas and discussions, multiple novelties and advantages of the current article are observed:

- The Current article deals with the detailed algorithmic descriptions of all the genres of image caption generation methods which are not explained in any of the papers published in recent times.
- The article showcases generic primitive and deep-neural-network-based algorithms which are addressed first time in any article related to image caption analysis or image caption generation.
- The article addresses all the recent highly peer-reviewed articles related to image caption analysis, which were published during the last decade.
- The article addresses almost all the datasets such as MS-COCO, Flickr, and others which are used by multiple innovative methods as mentioned in section 2.
- The article addresses almost all the performance parameters such as BELU, SPICE, and others that are used to measure the performances of any image caption generation method in either a quantitative or qualitative way.
- The article showcases the performance analysis of various recent methodologies using MS-COCO and Flickr datasets. As MS-COCO and Flickr are considered to be the most versatile, and bigger datasets currently available on the web, hence these datasets are considered in this article to evaluate the performances of various recent methodologies. As demonstrated in the comprehensive survey in Section 2, none of the recent studies utilize both the MS-COCO and Flickr datasets to evaluate image caption generation performance in both quantitative and qualitative manners.

7 Limitations of the Current Article

Despite multiple novelties and advantages, the article poses several limitations.

- In this article, MS-COCO, and Flickr datasets are considered only during the results and analysis phase to measure the perfection in execution of the recent image caption generation methods in both quantitative ways. The

article does not consider the rest of the datasets as mentioned in Section 4.2, due to their non-versatility, poor data formats, size, resolution, and numbers. In the future, these datasets may be tested on all the methods as mentioned in Resultant Tables I to IV.

- The Resultant Tables I to IV contain some non-computable values based on some performance parameters and existing methodologies. In the future, these non-computable values may be measured for further analysis.
- In the future, more innovative and high-performing algorithms should be developed based on the detailed analysis presented in this article. So, that a state-of-art performance can be achieved.

8 Conclusion

This article conducts a comprehensive analysis of image captioning techniques developed and employed in the past decade. It begins by examining diverse methodologies, explaining generic algorithms, introducing unique caption generation algorithms, and exploring their applications, strengths, and weaknesses. Subsequently, it compares these methods in terms of their utilization of different datasets and various performance metrics. Finally, the article delves into the results, scrutinizing them across various parameters. This study offers a thorough structural examination of recent image captioning techniques, elucidating the progression of image caption generation over the past few decades for researchers. The insights from this analysis will be invaluable for future research, potentially leading to the development of a robust technique capable of helping with the mentioned constraints of existing methods while retaining their valuable features.

Data Availability

The data used to hold up the findings of this study are present within the article itself.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding Statement

This study did not receive any funding.

References

- [1] Wang, H., Zhang, Y., Yu, X.: An overview of image caption generation methods. *Computational Intelligence and Neuroscience* **2020**, 1–13 (2020). <https://doi.org/10.1155/2020/3062706>
- [2] Ye, S., Han, J., Liu, N.: Attentive linear transformation for image captioning. *IEEE Transactions on Image Processing* **27**(11), 5514–5524 (2018). <https://doi.org/10.1109/TIP.2018.2855406>
- [3] Lu, X., Wang, B., Zheng, X., Li, X.: Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing* **56**(4), 2183–2195 (2018). <https://doi.org/10.1109/TGRS.2017.2776321>
- [4] Yu, N., Hu, X., Song, B., Yang, J., Zhang, J.: Topic-oriented image captioning based on order-embedding. *IEEE Transactions on Image Processing* **28**(6), 2743–2754 (2019). <https://doi.org/10.1109/TIP.2018.2889922>
- [5] Yang, M., Zhao, W., Xu, W., Feng, Y., Zhao, Z., Chen, X., Lei, K.: Multitask learning for cross-domain image captioning. *IEEE Transactions on Multimedia* **21**(4), 1047–1061 (2019). <https://doi.org/10.1109/TMM.2018.2869276>
- [6] Park, C.C., Kim, B., KIM, G.: Towards personalized image captioning via multimodal memory networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **41**(4), 999–1012 (2019). <https://doi.org/10.1109/TPAMI.2018.2824816>
- [7] Pham, P.T., Moens, M.-F., Tuytelaars, T.: Cross-media alignment of names and faces. *IEEE Transactions on Multimedia* **12**(1), 13–27 (2010). <https://doi.org/10.1109/TMM.2009.2036232>
- [8] Cheng, L., Wei, W., Mao, X., Liu, Y., Miao, C.: Stack-vs: Stacked visual-semantic attention for image caption generation. *IEEE Access* **8**, 154953–154965 (2020). <https://doi.org/10.1109/ACCESS.2020.3018752>
- [9] Huang, Y., Chen, J., Ouyang, W., Wan, W., Xue, Y.: Image captioning with end-to-end attribute detection and subsequent attributes prediction. *IEEE Transactions on Image Processing* **29**, 4013–4026 (2020). <https://doi.org/10.1109/TIP.2020.2969330>
- [10] Wang, B., Wang, C., Zhang, Q., Su, Y., Wang, Y., Xu, Y.: Cross-lingual image caption generation based on visual attention model. *IEEE Access* **8**, 104543–104554 (2020). <https://doi.org/10.1109/ACCESS.2020.2999568>

- [11] Xu, N., Zhang, H., Liu, A.-A., Nie, W., Su, Y., Nie, J., Zhang, Y.: Multi-level policy and reward-based deep reinforcement learning framework for image captioning. *IEEE Transactions on Multimedia* **22**(5), 1372–1383 (2020). <https://doi.org/10.1109/TMM.2019.2941820>
- [12] Ji, J., Xu, C., Zhang, X., Wang, B., Song, X.: Spatio-temporal memory attention for image captioning. *IEEE Transactions on Image Processing* **29**, 7615–7628 (2020). <https://doi.org/10.1109/TIP.2020.3004729>
- [13] Wu, J., Chen, T., Wu, H., Yang, Z., Luo, G., Lin, L.: Fine-grained image captioning with global-local discriminative objective. *IEEE Transactions on Multimedia* **23**, 2413–2427 (2021). <https://doi.org/10.1109/TMM.2020.3011317>
- [14] Zhang, S., Zhang, Y., Chen, Z., Li, Z.: Vsam-based visual keyword generation for image caption. *IEEE Access* **9**, 27638–27649 (2021). <https://doi.org/10.1109/ACCESS.2021.3058425>
- [15] Hou, D., Zhao, Z., Liu, Y., Chang, F., Hu, S.: Automatic report generation for chest x-ray images via adversarial reinforcement learning. *IEEE Access* **9**, 21236–21250 (2021). <https://doi.org/10.1109/ACCESS.2021.3056175>
- [16] Zhou, Z., Xu, L., Wang, C., Xie, W., Wang, S., Ge, S., Zhang, Y.: An image captioning model based on bidirectional depth residuals and its application. *IEEE Access* **9**, 25360–25370 (2021). <https://doi.org/10.1109/ACCESS.2021.3057091>
- [17] Wu, L., Xu, M., Sang, L., Yao, T., Mei, T.: Noise augmented double-stream graph convolutional networks for image captioning. *IEEE Transactions on Circuits and Systems for Video Technology* **31**(8), 3118–3127 (2021). <https://doi.org/10.1109/TCSVT.2020.3036860>
- [18] Liu, M., Hu, H., Li, L., Yu, Y., Guan, W.: Chinese image caption generation via visual attention and topic modeling. *IEEE Transactions on Cybernetics* **52**(2), 1247–1257 (2022). <https://doi.org/10.1109/TCYB.2020.2997034>
- [19] Ben, H., Pan, Y., Li, Y., Yao, T., Hong, R., Wang, M., Mei, T.: Unpaired image captioning with semantic-constrained self-learning. *IEEE Transactions on Multimedia* **24**, 904–916 (2022). <https://doi.org/10.1109/TMM.2021.3060948>
- [20] Zha, Z.-J., Liu, D., Zhang, H., Zhang, Y., Wu, F.: Context-aware visual policy network for fine-grained image captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**(2), 710–722 (2022). <https://doi.org/10.1109/TPAMI.2019.2909864>

- [21] Mahadi, M.R.S., Arifianto, A., Ramadhani, K.N.: Adaptive attention generation for indonesian image captioning. In: 2020 8th International Conference on Information and Communication Technology (ICoICT), pp. 1–6 (2020). <https://doi.org/10.1109/ICoICT49345.2020.9166244>
- [22] Tariq, A., Foroosh, H.: A context-driven extractive framework for generating realistic image descriptions. *IEEE Transactions on Image Processing* **26**(2), 619–632 (2017). <https://doi.org/10.1109/TIP.2016.2628585>
- [23] Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **39**(4), 664–676 (2017). <https://doi.org/10.1109/TPAMI.2016.2598339>
- [24] Javed, A., Bajwa, K.B., Malik, H., Irtaza, A.: An efficient framework for automatic highlights generation from sports videos. *IEEE Signal Processing Letters* **23**(7), 954–958 (2016). <https://doi.org/10.1109/LSP.2016.2573042>
- [25] Yanagimoto, H., Shozu, M.: Multiple perspective caption generation with attention mechanism. In: 2020 9th International Congress on Advanced Applied Informatics (IIAI-AAI), pp. 110–115 (2020). <https://doi.org/10.1109/IIAI-AAI50415.2020.00031>
- [26] Tiwary, T., Mahapatra, R.: An accurate generation of image captions for blind people using extended convolutional atom neural network. *Multi-media Tools and Applications* **82**, 1–30 (2022). <https://doi.org/10.1007/s11042-022-13443-5>
- [27] Ghandi, T., Pourreza, H., Mahyar, H.: Deep learning approaches on image captioning: A review. *ACM Computing Surveys* (2023). <https://doi.org/10.1145/3617592>
- [28] Luo, G., Cheng, L., Jing, C., Zhao, C., Song, G.: A thorough review of models, evaluation metrics, and datasets on image captioning. *IET Image Processing* **16** (2021). <https://doi.org/10.1049/ipr2.12367>
- [29] Roy, A.: A Guide to Image Captioning. Accessed on August 31, 2023 (2020). <https://towardsdatascience.com/a-guide-to-image-captioning-e9fd5517f350>
- [30] Tran, K., He, X., Zhang, L., Sun, J., Carapcea, C., Thrasher, C., Buehler, C., Sienkiewicz, C.: Rich image captioning in the wild (2016)
- [31] Kameswari, A.: Image caption generator using deep learning. *International Journal for Research in Applied Science and Engineering Technology* **9**, 1554–1564 (2021). <https://doi.org/10.22214/ijraset.2021.38652>

- [32] Al-Malla, M., Jafar, A., Ghneim, N.: Image captioning model using attention and object features to mimic human image understanding. *Journal of Big Data* **9**, 20 (2022). <https://doi.org/10.1186/s40537-022-00571-w>
- [33] Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. *ArXiv* **1409** (2014)
- [34] Hui, J.: Soft hard attention. Accessed on August 31, 2023 (2017). <https://jhui.github.io/2017/03/15/Soft-and-hard-attention/>
- [35] Doshi, K.: Transformers Explained Visually (Part 3): Multi-head Attention, deep dive. Accessed on August 31, 2023 (2021). <https://towardsdatascience.com/transformers-explained-visually-part-3-multi-head-attention-deep-dive-1c1ff1024853>
- [36] Luong, M.-T., Pham, H., Manning, C.: Effective approaches to attention-based neural machine translation (2015). <https://doi.org/10.18653/v1/D15-1166>
- [37] Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing when to look: Adaptive attention via a visual sentinel for image captioning (2016)
- [38] You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4651–4659 (2016). <https://doi.org/10.1109/CVPR.2016.503>
- [39] Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., Chua, T.-S.: Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 6298–6306 (2016)
- [40] Pedersoli, M., Lucas, T., Schmid, C., Verbeek, J.: Areas of attention for image captioning, pp. 1251–1259 (2017). <https://doi.org/10.1109/ICCV.2017.140>
- [41] Dan, Z., Fang, Y.: Deliberate Multi-Attention Network for Image Captioning, pp. 475–487 (2022). https://doi.org/10.1007/978-3-031-18907-4_37
- [42] Guo, L., Liu, J., Yao, P., Li, J., Lu, H.: Mscap: Multi-style image captioning with unpaired stylized text. *Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2019)

- [43] Renato Sortino, F.R. Simone Palazzo, Spampinato, C.: Transformer-based image generation from scene graphs. *Computer Vision and Image Understanding*, 103721 (2023). <https://doi.org/10.1016/j.cviu.2023.103721>
- [44] Renato Sortino, S.P., Spampinato, C.: Transformer-Based Scene Graph to Image. Accessed on August 31, 2023 (2023). <https://github.com/perceivelab/trf-sg2im/blob/main/images/architecture.png>
- [45] Chen, H., Wang, Y., Yang, X., Li, J.: Captioning transformer with scene graph guiding. In: 2021 IEEE International Conference on Image Processing (ICIP), pp. 2538–2542 (2021). <https://doi.org/10.1109/ICIP42928.2021.9506193>
- [46] Pan, J.-Y., Yang, H.-j., Faloutsos, C., Duygulu, P.: Gcap: Graph-based automatic image captioning (2004). <https://doi.org/10.1109/CVPR.2004.79>
- [47] Zhong, Y., Wang, L., Chen, J., Yu, D., Li, Y.: Comprehensive image captioning via scene graph decomposition. In: ECCV (2020)
- [48] Wang, Q., Chan, A.: CNN+CNN: Convolutional Decoders for Image Captioning
- [49] Xu, H., Yan, M., Li, C., Bi, B., Huang, S., Xiao, W., Huang, F.: E2E-VLP: End-to-End Vision-Language Pre-training Enhanced by Visual Learning
- [50] Gautam, T.: Implementation of Attention Mechanism for Caption Generation on Transformers using TensorFlow. Accessed on August 31, 2023 (2021). <https://www.analyticsvidhya.com/blog/2021/01/implementation-of-attention-mechanism-for-caption-generation-on-transformers-using-tensorflow/>
- [51] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.: Microsoft coco: Common objects in context, vol. 8693 (2014). https://doi.org/10.1007/978-3-319-10602-1_48
- [52] Hsankesara: Flickr Image Dataset. Accessed on August 31, 2023 (2018). <https://www.kaggle.com/datasets/hsankesara/flickr-image-dataset>
- [53] Lu, X., Wang, B., Zheng, X., Li, X.: Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing* **56**(4), 2183–2195. <https://doi.org/10.1109/TGRS.2017.2776321>
- [54] Yusnu, L.E.M.: Oxford 102 Flower Dataset. Accessed on August 31, 2023 (2021). <https://www.kaggle.com/datasets/nunenuh/pytorch-challenge-flower-dataset>

- [55] Park, C., Kim, B., Kim, G.: Towards personalized image captioning via multimodal memory networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PP**, 1–1 (2018). <https://doi.org/10.1109/TPAMI.2018.2824816>
- [56] Thomee, B., Elizalde, B., Shamma, D., Ni, K., Friedland, G., Poland, D., Borth, D., Li, L.-J.: Yfcc100m: the new data in multimedia research. *Communications of the ACM* **59**, 64–73 (2016). <https://doi.org/10.1145/2812802>
- [57] Krause, J., Johnson, J., Krishna, R., Fei-Fei, L.: A hierarchical approach for generating descriptive image paragraphs (2016)
- [58] Raddar: Chest X-rays (Indiana University). Accessed on August 31, 2023 (2020). <https://www.kaggle.com/datasets/raddar/chest-xrays-indiana-university>
- [59] Johnson, P.T.M.R.B.S..H.S. A.: MIMIC-CXR Database (version 2.0.0). PhysioNet. Accessed on August 31, 2023 (2019). <https://doi.org/10.13026/C2JT1Q>
- [60] Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst (October 2007)
- [61] Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation (2002). <https://doi.org/10.3115/1073083.1073135>
- [62] Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments (2005)
- [63] Lin, C.-Y.: Rouge: a package for automatic evaluation of summaries (2004)
- [64] Vedantam, R., Zitnick, C.L., Parikh, D.: Cider: Consensus-based image description evaluation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4566–4575 (2015). <https://doi.org/10.1109/CVPR.2015.7299087>
- [65] Dalianis, H.: *Evaluation Metrics and Evaluation*, pp. 45–53. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-78503-5_6. https://doi.org/10.1007/978-3-319-78503-5_6
- [66] wikipedia.org: Cross entropy. Accessed on August 31, 2023 (2023). https://en.wikipedia.org/wiki/Cross_entropy/

- [67] Chapter 5 - fuzzy measures. probabilities/possibilities. In: Dubois, D., Prade, H. (eds.) Fuzzy Sets and Systems. Mathematics in Science and Engineering, vol. 144, pp. 125–147. Elsevier (1980). [https://doi.org/10.1016/S0076-5392\(09\)60141-7](https://doi.org/10.1016/S0076-5392(09)60141-7). <https://www.sciencedirect.com/science/article/pii/S0076539209601417>
- [68] dimensions.freshdesk.com: What is "Relevance" and how is it calculated? Accessed on August 31, 2023 (2021). <https://dimensions.freshdesk.com/support/solutions/articles/23000022475/>