International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

A Quantitative Based Research on the Production of Image Captioning

Samuel-Soma M. Ajibade*[1], Abdelhamid Zaidi[2], Siti Sarah Maidin[3], Wan Hussain Wan Ishak*[4], Adedotun Adetunla[5]

Submitted: 07/05/2023 **Revised**: 12/07/2023 **Accepted**: 06/08/2023

Abstract: It is widely recognized that modern systems can discern the context of an image and enrich it with relevant captions through the fusion of computer vision and natural language processing, a technique referred to as 'image caption production.' This article aims to shed light on and analyze various image captioning techniques that have evolved over the past few decades, including the Attention Model, Region-Level Caption Detection, Semantic Content-Based Models, Multimodal Models, and more. The evaluation of these techniques employs diverse criteria such as Precision Rate, Recall Rate, F1 Score, Accuracy Rate, among others, while employing various datasets for comparison. This article offers a comprehensive structural examination of contemporary image captioning methods. Researchers can leverage the insights from this analysis to develop innovative, improved approaches that sidestep the shortcomings of older methods while retaining their beneficial aspects.

Keywords: *Attention Model, Image Caption, Multimodal Model, Region Level Captions, Semantic Content*

1. Introduction

Image caption production is a process of recognizing the image context and annotating the image context with relevant captions based on computer-based techniques. In this technique, an image is labelled with suitable keywords. During the model training, we use lots of datasets which is helpful in context labelling. The caption generation from images is considered to be a tough, complex, and difficult task. The main challenge of this task is to read the different objects in the image and to extract a meaningful caption in a natural language (like English). Emulating the human capacity of giving captions to images by a machine is itself an exceptional task along the line of Artificial Intelligence. Though the task is significantly harder it would create a great impact on visually impaired people to understand the content of images. A description not only gathers the information about the objects contained in an image but also communicates how those objects are related to each other as well as the activities they are engaged in. The above procedure needs to be implemented in a natural language like English, in which a language model is required. Traditionally, computer systems have been utilizing pre-defined formats for creating text descriptions for images. However, this method does not give sufficient variety required for producing lexically rich text descriptions. Along with the evolution of ANN, such disadvantages are overcome.

It is the task of an intelligence system to identify, recognize, understand, and detect captions from a digital image. An image caption generation model is constituted by the discovery of objects from a digital image and revealing their connection with a natural language like English. Some of the recent works also revealed that an attention-based model is also helpful to store the context of an image and successful to establish a relationship between clusters and silent features of an image.

Image captioning is used in various cases such as helping a visually impaired individual using text to speech by real-time responses about the surrounding environment through a camera, enhancing social media platforms by changing captions for images in social feed additionally as messages to speech, helping young kids in recognizing objects as well as learning the English language. Captions for every image on the internet will result in quicker and descriptively accurate image searches and indexing. In robotics, the information gathered by an agent about the surrounding environment can be given a context through natural language representation of surroundings through the captions for the images in the camera feed. There are few real-world applications where this approach can be very useful. In the case of Self-driving vehicles, automatic driving is the greatest challenge and if we can properly caption the scene around the vehicle then it can open a brand-new possibility in the self-driving system. We can make a product for the blind which will guide them while they are walking on the roads without the support of anyone else. We can do this by first capturing the surrounding environment of the person with the help of a camera feed

[1**]*Department of Computer Engineering, Istanbul Ticaret University, Turkiye*
[2] *Department of Mathematics, College of Science, Qassim University, P.O. Box 6644, Buraydah 51452, Saudi Arabia*
[3] *Faculty of Data Science & Information Technology, INTI International University, Nilia, Malaysia*
[4**] *School of Computing, Universiti Utara Malaysia, Malaysia*
[5]*Department of Mechanical and Mechatronics Engineering, Afe Babalola University, Ado Ekiti, Nigeria*
** Corresponding Author Email: asamuel@ticaret.edu.tr, hussain@uum.edu.my*

and converting it into text and then the text to voice. If we can generate relevant captions about the scenes captured by the CCTV cameras, then whenever a malicious activity goes on somewhere it will automatically raise alarms. This might probably help reduce some crime and/or accidents. Automatic Captioning will help to make a Google Image Search pretty much as good as Google Search and then every image could be first converted into a caption and so a search can be performed based on the caption.

During the last few decades, various image captioning techniques are proposed by researchers. These techniques are well appreciated by the scientific societies, but the demand of the society is not fulfilled. As a result, to meet the demand of society, various Image caption generation techniques are evolved. Hence, an organized analysis of the recent techniques in this domain is absolutely necessary to evolutes their pros and cons and performance. This article highlights an analysis of different methodologies concerning their applications, algorithms, and merits and demerits. The structure of this paper is represented by multiple sections. In this paper in section 2, we are presenting some of the related works associated with the image caption generation from the last decades. We are trying to highlight the merits and demerits of various image caption methodologies, significant descriptions of different image caption generation datasets, and different parameters for measuring the performance of various image caption methodologies. In section 3, we are trying to highlight some generic structures for image captioning by providing their definition, algorithms, working principles, and architectures. In section 4, we are trying to highlight the values received for different parameters for different methods as described in section 2. We are presenting all these results with the help of several resultant tables to analyze which method is superior in terms of a parameter to the other.

2. Literature Review

In the domain of image caption generation, many techniques are proposed over the last few decades. Attentive linear transformation [2] [49] performs the linear transformation weights containing valuable information without any kind of concrete form. High performance is achieved with the help of this technique. This model can be applied in visual question answering and neural machine translation. While the model failed to identify words on the sign in some cases. The model failed to segregate intersecting objects in some cases. The model failed to calculate the right quantity in some cases. The model is failed to identify the gender of a person in some cases. A Multimodal Method [7] method was proposed which includes RNN [48], LSTM Attention-Based Method. It is capable of describing an RSICD dataset for remote sensing images by considering some special characteristics. A huge number of remote sensing images is produced in the dataset which is very rare to be found in any remote sensing device. Calculations of performance based on popular remote sensing methods are conducted for larger analysis. But prepared dataset used by this method can be enhanced whereas some sentences are produced by overwriting the previous sentences from the remote sensing images and no novel image caption generation model is introduced. A Topic oriented model [9] gives satisfactory results w.r.t. considered MSCOCO dataset, while the quality of the caption is very high. But the result of this method w.r.t. the other dataset is not produced.

The Multitask Learning Algorithm for Cross-Domain Image Captioning [12] gives high performance but fails to distinguish images with visual symmetry. It provides a low-performance score for random samples. The Context Sequence Memory Network (CSMN) [3] model suggested that the multiple types of context information can be stored from the image. The algorithm is useful to store long-term information and context understanding. It failed to include various types of metadata. There is a scope of post commenting. The method does not apply to all social media. The Unsupervised Cross-Media Alignment [1] model can perform alignment of phases, and word text conversion, and supports multiple languages. But it suffers from a lack of accuracy. Stack-VS [13] is a stack decoder-based model that generates visual and semantic level caption, and fine-grained caption. It is observed that a reasonable caption is not generated through this method. The model may include a graph convolutional model for better performance.Multimodal attribute detector and subsequent attribute predictor [14] model is capable of dynamic attribute prediction, and precise caption generation. But the model failed to produce relational attributes.The Visual Attention Model [15] is a cross-lingual model, which is an independent recurrent structure. It is capable of performing feature and semantic similarity analysis. But failed to produce language-specific tasks and suffers from information loss.The multi-level policy and reward RL framework [10] model is capable to perform word and sentence level caption generation, vision-language, and language-language reward. We need to train the policy network to generate the output.The spatio temporal memory attention [16] model produces a strong temporal connection between attention and good performance. It is used to learn the Spatio-temporal relationship of attended areas. It is observed that it is not a generic model. The global-local discriminative objective [17] model enhances discriminability. It is capable of productive image caption generation and fine-grained caption generation. In this method, the local discriminative threshold value may not be adjusted. Its discriminative objective overwrites frequently used words. Visual Semantic Attention Model (VSAM) [20] visual keyword concept generates precise and valuable captions and it is effective for visual keyword extraction.

It is found that it is not a fully developed image caption framework and the precision rate can be improved. Adversarial reinforced report-generation framework [21] is a novel XRay caption generation framework that is not capable of removing noise from the X-Ray images. Bidirectional depth residuals gated recurrent unit network [22] gives a high prediction rate, and low inference time. But it suffers from poor model stability. Noise Augmented Double-stream Graph Convolutional Networks (NADGCN) [18] model is capable of extracting full image context. It is also capable to extract context from the background. It has a low vernal ability. It suffers from architectural complexity.NICVATP2L [19] model produces low accuracy in language generation and shows low diversity in multiattribute entities. But it is capable of generating descriptive and informative captions. Semantic-Constrained Self-learning (SCS) [23] model is capable of effective semantic object detection and produces state-of-the-art unpaired captioning. But it is expensive and requires a complex experimental setup. Context-Aware Visual Policy network (CAVP) [11] model is capable of efficient sentence captioning and produces improved paragraph captioning. But there is a scope for improvement in sentence and paragraph captioning, performance matrices can be improved, and there is a scope to improve sequential decision-making operations. Task-Adaptive Attention module [4] is a non-visual features extraction technique, and expression to caption conversion. There is a scope for performance improvement. The method may apply to attention-based encoder-decoder image captioning.

Context-Driven Extractive Method [8] 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 [5] is capable of generating a description using one input array, but it has a significant drawback to produce region-level captions. Gradual Transition and Scope-Caption Detection [6] is a computationally efficient robust methodology. The authors failed to produce the application of this method on a large size dataset. An Attention Mechanism [24] is proposed to generate and control image captions. But the method suffers from a lack of resultant table and requires lots of performance metrics to determine the original result.

2.1. Datasets

2.1.1. MS-COCO

This dataset is [37] developed by Microsoft Corporation for the detection of a large-scale object, for the segmentation of objects, and to perform image caption. The dataset has a feature for object segmentation with detailed instance annotations. The dataset has other features such as superpixel stuff segmentation. It contains over 330000 labeled images. It contains 1.5M object instances and contains 5 captions per image. Multiple methods have used this dataset for their parametric calculations.

2.1.2. Flickr

The Flickr dataset [38] is a large dataset that contains over 31,000 images. Multiple methods have used this dataset for their parametric calculations.

2.1.3. The Remote Sensing Image Captioning Dataset (RSICD)

This dataset [39] is capable of producing image captions from over 10k remote sensing images. Images in this dataset are collected from various sources like Google Earth. The size of each image is 224 pixels x 224 pixels. All the images present in this dataset may vary. An image in this dataset is capable of producing descriptions of 5 sentences.

2.1.4. Oxford 102

It is a flower dataset [40]. This dataset is majorly used for image classification. There are 102 categories of flower images in this dataset. The flower used in this dataset is collected from different parts of Europe. Every class of flowers contains a minimum of 40 images and a maximum of 258 images.

2.1.5. InstaPIC-1.1M

This is a publicly available Kaggle dataset [3].

2.1.6. YFCC100M

It is an image and video dataset [41] with a 100 million capacity.

2.1.7. Stanford Image Paragraph Captioning

It is generated from another dataset [42] which is known as the Visual Genome dataset. This is a large dataset that contains approximately 20,000 images. The images in this dataset are labeled with their corresponding paragraphs. It is useful for captioning a paragraph.

2.1.8. IU X-Ray

IU X-Ray is an open access chest X-ray dataset [43] from Indiana University. There is a total of 7466 images in this dataset.

2.1.9. MIMIC-CXR

It is a large public dataset [44]. It contains radiology images (CT/MRI) of the chest. All the images are in DICOM format. The images are completely free of text associated with radiological reports. It holds about 377,110 images from which 227,835 images are collected from various radiographic studies. The dataset is collected from the studies, performed at the Beth Israel Deaconess Medical Center in Boston, MA.

2.1.10. IVKD

IVKD is an image Visual Keyword dataset [20]. In the

context of image caption analysis, it is observed that it is used to generate multiple captions.

2.1.11. Chinese Image Caption (AIC-ICC)

This dataset [19] is used for the generation of Chinese image captions. It is considered to be the largest dataset for generating the Chinese image caption.

2.1.12. Labeled Faces in the Wild (LFW)

This dataset [45] is used for unconstrained face recognition. There are over 13K images present in this dataset form and more than 5K people's faces are detected using the Viola-Jones face detector.

2.1.13. News image dataset

This is a custom dataset [8] prepared from news broadcast videos. It is a private dataset.

2.1.14. Sports video dataset

This is a custom dataset [6] prepared from sports broadcast videos. It is a private dataset.

2.2. Performance Metrics

It is very important to measure the performance or accuracy of the image caption generation procedure after its completion. Most of the image captioning methods use BELU, METEOR, ROUGE, CIDER, and SPICE parameters to measure the performance of the image captioning. But some of the methods [1][6][8][14][21]are used to recall, precision, accuracy score, and F1 score to calculate the performance. It has been observed that some of the methods [25] measure their performances using cross-entropy. Some of the existing methodologies [3] introduce Plausibility, Grammaticality, and Relevance to measure the performances. These metrics can be calculated using some predetermined method. In the next section of this article different parameters are analyzed.

2.2.1. BLEU

The BLEU [26] stands for Best Linear Unbiased Estimator. In this parameter, the word best stands for minimum variance. The parameter can be calculated using the following equations.

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

Where,

$\beta \leftarrow$ The brevity penalty

$\mu \leftarrow$The length of the candidate translation

$\theta \leftarrow$The effective reference corpus length

$\varepsilon \leftarrow$ Theresiduals

$$B \leftarrow \beta.\exp\left(\sum_{i=1}^{K} \omega_i \log\rho_i\right) \quad (2)$$

Where,

$\beta \leftarrow$ The brevity penalty

exp ← The exponential

$\rho_i \leftarrow$The i-gram precisions

$\omega_i \leftarrow$ Theithpositiveweight

i← ithgram

K ← Kth gram

B ← BLEUParameter

2.2.2. METEOR

The METEOR [27] metric is used to overcome the drawbacks of the BLEU metric. The metric is calculated in the following way:

$$f_\alpha \leftarrow \frac{10rp}{r+9p} \quad (3)$$

Where,

$f_\alpha \leftarrow$ The functional mean

r ← The unigram recall

P ← unigram precision

r + 9p ← The harmonic-mean

If,

c ← No of possible chunks

n ← No of the unigrams matched

P ← Penalty

Then, P can be calculated as a

$$P\,P \leftarrow 0.5 * \left(\frac{c}{n}\right) \quad (4)$$

With the help of the functional mean (f_α) and penalty (P) we can calculate the METEOR Score (M_{Score}) for the given alignment.

$$M_{Score} \leftarrow f_\alpha*(1-P) \quad (5)$$

2.2.3. Rouge

This parameter [28] is used to calculate text summaries. This parameter can be calculated as rouge − i, rouge − land rouge − s. rouge − i is a parameter that is denoted by i-gram recall between the reference and the candidate. rouge − i is calculated as follows:

$$\text{rouge} - \text{i} \leftarrow \frac{\sum_{\varepsilon\in\{\text{Ref Summaries}\}}\sum_{g_i\in\varepsilon} C_m(g_i)}{\sum_{\varepsilon\in\{\text{Ref Summaries}\}}\sum_{g_i\in\varepsilon} C(g_i)} \quad (6)$$

Where,

rouge − n ← Rouge parameter for n − gram recall

$i \leftarrow$the length of the i-gram

$g_i \leftarrow$ ith gram

$C_m(g_i) \leftarrow$ max no of i $-$ grams in the candiadte summaries

$\varepsilon \leftarrow$ The ref summaries set

$\text{rouge} - \text{lis}$ a parameter that is denoted by LCS-based data. $\text{rouge} - \text{l}$ is calculated as follows:

$$\text{rouge} - \text{l}(c, r) \leftarrow (1 + \delta^2) r^{lcs}(c, r) p^{lcs}(c, r) r^{lcs}(c, r) + \delta^2 p^{lcs}(c, r) \quad \textbf{(7)}$$

Where,

$\delta \leftarrow$ the relative importance of the precision and recall

$\text{rouge} - \text{l}(c, r) \leftarrow$ the Metric between a candidate document and a single reference document

$c \leftarrow$a candidate document

$r \leftarrow$ a single reference document

$r^{lcs}(c, r) \leftarrow$ the recall score of the set of longest common subsequences in the candidate document c and the reference document r.

$p^{lcs}(c, r) \leftarrow$ the precision score of the set of longest common subsequence in the candidate document c and the reference document r.

$\text{rouge} - \text{s}$ is a parameter that is denoted by i-gram with skips.$\text{rouge} - \text{s}$ is calculated as follows:

$$\text{rouge} - \text{s}(c, r) \leftarrow \frac{(1+\delta^2) r_s(c,r) p_s(c,r)}{r_s(c,r) + (\delta^2) p_s(c,r)} \quad \textbf{(8)}$$

Where,

$\delta \leftarrow$ The relative importance of the precision and recall

$\text{rouge} - \text{s}(c, r) \leftarrow$ the Metric for an F-score measure between a candidate document and a single reference document

$c \leftarrow$a candidate document

$r \leftarrow$ a single reference document

$r_s(c, r) \leftarrow$ the recall score of the set of skip-bigram in the candidate document c and the reference document r.

$p_s(c, r) \leftarrow$ the precision score of the set of skip-bigram in the candidate document c and the reference document r.

2.2.4. Cider

This parameter [30] can be calculated using the following formula:

$$\text{cider} - \text{param}(c_m, S_m) \leftarrow \frac{1}{k} \sum n \frac{f^l(c_m).f^l(s_{mn})}{||f^l(c_m)||.||f^l(s_{mn})||}$$

(9)

Where,

$\text{cider} - \text{param}(c_m, S_m) \leftarrow$A score for l-grams of length l is computed using the average cosine similarity between the candidate sentence and the reference sentences.

$f^l(c_m), f^l(s_{mn}) \leftarrow$are vectors.

$||f^l(c_m)||, ||f^l(s_{mn})| \leftarrow$the magnitude of the vectors

2.2.5. Spice

This parameter [29] can be calculated using the following formulas:

$$p(c, R) \leftarrow t(c) \theta t(R) / t(c) \quad (10)$$

$$r(c, R) \leftarrow t(c) \theta t(R) / t(R) \quad (11)$$

$$\text{spice} - \text{param}(c, R) \leftarrow 2 * p(c, R) * r(c, R) / p(c, R) * r(c, R) \quad (12)$$

Where,

$c \leftarrow$the caption for a candidate

$R \leftarrow$A reference caption set

$t \leftarrow$the caption to the tuple mapping function

$\text{spice} - \text{param}(c, R) \leftarrow$ The spice parameter with argument c and R

2.2.6. Precision Rate

It can be calculated [31] as follows:

$$\text{precision} - \text{rate} \leftarrow tp / tp + fp \quad (13)$$

Where,

$tp \leftarrow$the no of true positive samples

$fp \leftarrow$ the no of false-positive samples

2.2.7. Recall Rate

It can be calculated [31] as follows:

$$\text{recall} - \text{rate} \leftarrow tp / tp + fn \quad (14)$$

Where,

$tp \leftarrow$the no of true positive samples

$fn \leftarrow$ the no of false-negative samples

2.2.8. F1 Score

It can be calculated [31] as follows:

$$f1 - \text{score} \leftarrow 2 * (p * r) / p + r \quad (15)$$

Where,

$p \leftarrow$the precision rate

$r \leftarrow$ the recall rate

2.2.9. Accuracy Score

It can be calculated [31] as follows:

$$\text{accuracy} - \text{score} \leftarrow \text{total no of correct prediction} / \text{total no of prediction} \quad (16)$$

2.2.10. Cross Entropy Loss

It can be calculated [32] as follows:

$$h(\alpha, \beta) \leftarrow \sum_{i \in I} \alpha(i) * \log\big(\beta(i)\big) \quad (17)$$

Where,

$h(\alpha, \beta) \leftarrow$the cross-entropy loss of the probabilities of the events from α and β

$\alpha(i) \leftarrow \alpha$ (i) is the probability of the event i in α

$\beta(i) \leftarrow$the probability of event i in β

The log is the base-2 logarithm.

2.2.11. Plausibility

It can be calculated [33] as follows:

(18)

Where,

$pl() \leftarrow$the plausibility function.

$\beta() \leftarrow$ the belief function

$\alpha \leftarrow$A subset α of a finite set X

$\overline{\alpha} \leftarrow$ complement of α

2.2.12. Relevance Score

The relevance score [34] is very important to calculate the image caption generation. The relevance in any dimension is a process by which a system can search for a keyword in the entire collection of written texts and assign the relative scores to those results having a successful match with the searched keywords. This relevance score is constituted based on a set of criteria. Some of the criteria are as follows:

- No occurrences of the keyword matching event in the entire text
- Is the keyword is found in the title of the entire text
- Is the keyword is found in the abstract of the entire text
- Is the keyword is found in the title and abstract of the entire text

The relevance score is a value. If the relevance score is high then it indicates that the searched keyword is more relevant to the result.

In the next section, Table. 1 Comparative Table - I is displayed which shows different methodologies suggested by different papers, the dataset used in these papers, and parameters to calculate the performance.

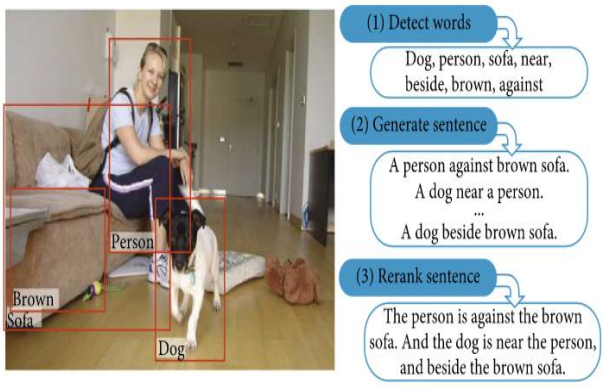

Fig. 1. Shows an image caption generation process.

3. Generic Structures for Image Captioning

In this section, we are explaining a generic structure of an image captioning method, after analyzing multiple methods displayed in the Comparative Table – I, which will be helpful to identify image captions with high accuracy.

3.1. Generic Model-I

It is an encoding and decoding framework for translating into a machine. This model is considered to be a generic model developed by Microsoft Corporation. In this architecture [25][35] four basic components are used. First, the feature extraction of an image is processed by a deep ResNet model [47]. Secondly, an image caption ranking and candidate generation language model is used. Third, entity recognition is used for the detection of celebrities and landmarks. The entity recognition detects if the image is of some celebrity or landmark. And fourth, the confidence score is estimated using a classifier. The confidence score is helpful to calculate the dependency of the captions. Figure 1 shows the image caption generation generic model I.

3.1.1. Procedure for Generic Model-I

In this section, the procedure for the Generic Model-I is explained. The procedure is constituted by the following steps.

Step 1: Insert an image into the system.

Step 2: The input image is transmitted to Convolutional Networks and its output is transmitted to the next level.

Step 3: The next level is constituted by a visual component extractor, celebrity information extractor, landmark information extractor, and feature vector extractor. This feature extraction model is processed by a deep ResNet model.

Step 4: Output of the visual component extractor, celebrity information extractor, and landmark information extractor is inserted into the language-based model.

Step 5: The output of the feature vector extractor is inserted into a Deep Multimodal Similarity Model (DMSM).

Step 6: Output of the language-based model and the DMSM model is inserted into the Confidence Generation Model.

Step 7: Output of the Confidence Generation Model gives the output caption.

Figure 2 shows the architecture of the image caption generation Generic Model-I.

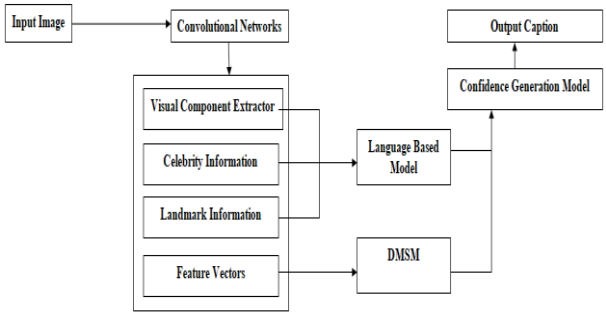

Fig. 3. Generic Model - I for image caption generation

3.2. Generic Model-II

This generic architecture [25][36], uses multiple Long Short-Term Memory (LSTM) modules. This model is considered to be a generic model developed by Google. This is an encoding and decoding model. In this model, the encoding operation is controlled by CNN. The image features are extracted using the last layer of the CNN [46]. The image description is generated using the decoding operation. The decoding operation is done by an LSTM model. The procedure of this generic model is explained below.

3.2.1. Procedure for Generic Model-II

In this section, the procedure for the Generic Model-II is explained. The procedure is constituted by the following steps:

Step 1: A feature vector of an image P is included only one time in the sequence of LSTM at an interval of -1.

Step 2: From time interval t = 0 to rest the whole word vector sequence is considered as input.

Step 3: The maximum probability of a step is calculated using an activation function that lies in the hidden state.

Step 4: The Beam Search technique is used for correct image caption prediction. At every time step t, it considers m no of sentences as best for a candidate prediction. For each time step t+1, the most probable words of m are considered along with m no of words and sentences. It generates m² no of probable sentences. At time step t+1, m best probable sentences are selected from the result.

Step 5: The final output word is generated. The equation responsible for generating the final output is calculated as follows:

$$\alpha_{-1} \leftarrow convolution(\rho) \quad (19)$$

$$\alpha_t \leftarrow \omega_\varepsilon \theta_t, t \in \{0,1,2,3,\dots,N-1\} \quad (20)$$

$$\beta_{t+1} \leftarrow lstm(\alpha_t), t \in \{0,1,2,3,\dots,N-1\} \quad (21)$$

Where,

$convolution(\rho)$ ←*A convolution function*

ρ ← *A feature vector of an image*

$t \in \{0,1,2,3,\dots,N-1\}$ ← *Every time step*

lstm ()← *A Long Short-Term Memory function*

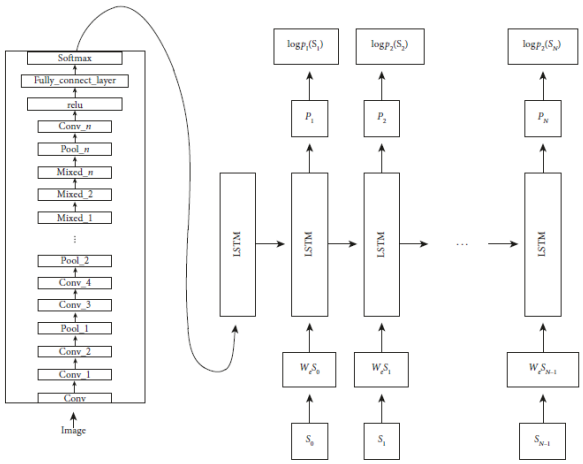

Fig. 3. Generic Model - II for image caption generation

4. Result and Discussion

This section represents a comparative study of various existing techniques based on different performance metrics as discussed in section 2.2. In this section first, multiple resultant tables are displayed which illustrate various existing methodologies proposed in the last few decades. The performances of these methodologies are calculated by various metrics such as BELU-1, BELU-2, BELU-3, BELU-4, METEOR, ROUGE, CIDER, and SPICE or their performances are calculated using Precession Rate, Recall Rate, F1 Score, and Accuracy Score. All the performances are calculated based on the MS-COCO datasets. After representing the performance value based on the above-mentioned parameters this section also highlights various graphs, plotted based on each parameter separately. Finally, different outputs in the form of images are projected together which are received after applying various existing techniques as mentioned earlier.

Resultant Table 1 highlights parameters such as BLEU-1, BELU-2, BELU-3, BELU-4, METEOR, ROUGE, CIDER, and SPICE to measure the performance of different existing methods. From the Resultant Table I it is clear that Multimodal Attribute Detector and Subsequent Attribute Predictor (MADASAP) [14] give the best performance in

terms of BELU-1, BELU-2, BELU-3, BELU-4, ROUGE, CIDER, and SPICE parameters, whereas it is observed that NICVATP2L [19] method overpowered rest of the method in terms of METEOR metric value. All results are calculated concerning the MS-COCO standard dataset and the process of normalization, balancing, and standardization are conducted on the dataset before the application of the dataset.

Resultant Table 2 highlights parameters such as Precession Rate, Recall Rate, F1 Score, and Accuracy Score to measure the performance of different existing methods whose performance cannot be measured using the parameters mentioned earlier in the Resultant Table 1. From the Resultant Table 2, it is clear that Gradual Transition and Scope-Caption Detection [6] give the best Precession Rate, Recall Rate, and Accuracy Rate whereas Unsupervised Cross-Media Alignment [1] gives the best F1 Score in comparison with other methods. We have calculated our results are calculated concerning the MS-COCO standard dataset and the process of normalization, balancing, and standardization are conducted on the dataset before the application of the dataset.

The hyphen symbol in the Resultant Table 1 and 2 indicates that the said parameter is not computed for the indicated method. Resultant Table II represents some of the methods which are non-computable for the parameters like BLEU, METEOR, ROUGE, CIDER, and SPICE. Resultant Table 1 represent some of the methods which are non-computable for the parameters like Precession Rate, Recall Rate, F1 Score, and Accuracy Score.

Figure 4 represents Graph-I for BLEU parameters [26]. Graph – I is projected to indicate different values of BELU parameters BELU-1, BELU-2, BELU-3, and BELU-4 – all these parameters are projected in a single graph. Their line of representation is highlighted in the graph with different color labels. The X-axis of the graph represents different image captioning techniques as discussed in the Resultant Table I. The Y-axis of the graph represents different BLEU parameter values. From this graph it can be concluded that the BELU-1, BELU-2, BELU-3, and BELU-4 parameters can give their highest values for MADASAP [14] techniques whereas MULTIMODAL [7] gives their lowest values for BLEU-1, BELU-2, BELU-3, and RGF [21] method gives the lowest value for BELU-4 respectively.

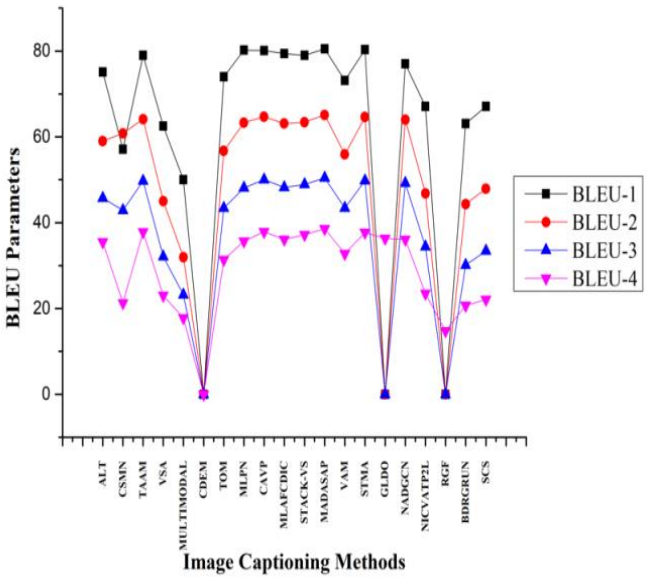

Fig. 4. Shows the Graph-I

Figure 5 represents Graph-II for METEOR [27] parameter representation. The X-axis of the graph represents different image captioning techniques as discussed in the Resultant Table I. The Y-axis of the graph represents different METEOR parameter values. From this graph, it can be concluded that the METEOR parameter can give its highest values for NICVATP2L [19] techniques whereas the ROUGE parameter gives its lowest value for VSA [5] method.

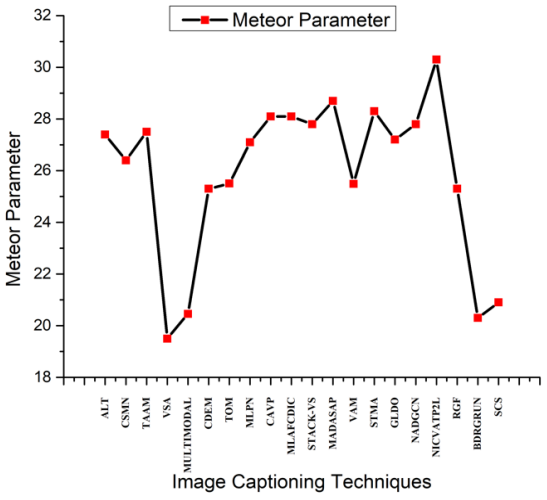

Fig. 5. Shows the Graph-II

Figure 6 represents Graph-III for ROUGE parameter representation. The X-axis of the graph represents different image captioning techniques as discussed in the Resultant Table I. The Y-axis of the graph represents different ROUGE parameter values. From this graph, it can be concluded that the ROUGE parameter can give its highest values for MADASAP [14] techniques whereas the ROUGE parameter gives its lowest value for ARRGF [21] method.

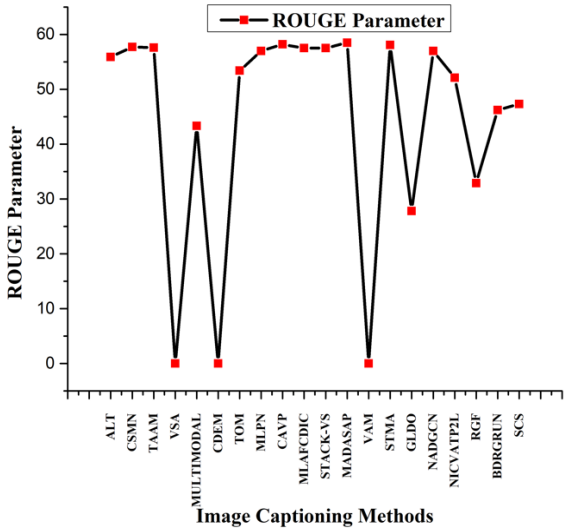

Fig. 6. Shows the Graph-III

Figure 7 represents Graph-IV for the CIDER parameter [30] representation. The X-axis of the graph represents different image captioning techniques as discussed in the Resultant Table I. The Y-axis of the graph represents different CIDER parameter values. From this graph, it can be concluded that the CIDER parameter can give its highest values for MADASAP [14] techniques whereas the CIDER parameter gives its lowest value for ARRGF [21] method.

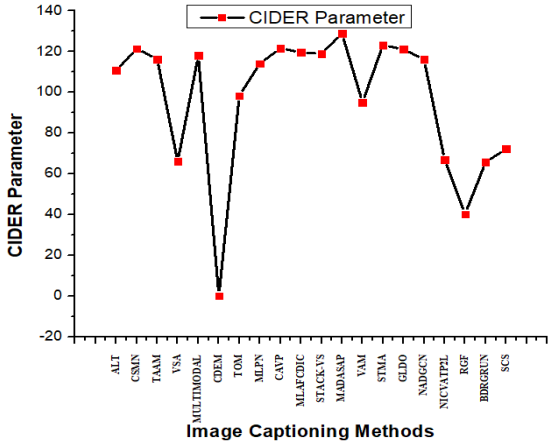

Fig. 7. Shows the Graph-IV

Figure 8 represents Graph-V for the SPICE parameter [29] representation. The X-axis of the graph represents different image captioning techniques as discussed in the Resultant Table I. The Y-axis of the graph represents different SPICE parameter values. From this graph, it can be concluded that the SPICE parameter [29] can give its highest values for MADASAP [14] techniques whereas the SPICE parameter gives its lowest value for BDRGRUN [22] method.

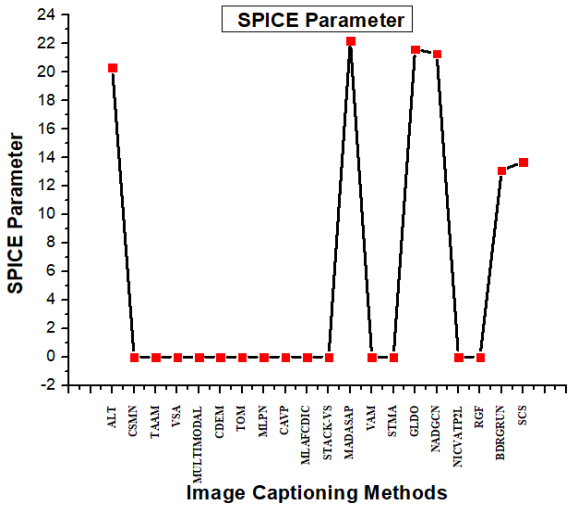

Fig. 8. Shows the Graph-V

Figure 9 represents Graph-VI for Precession Rate, Recall Rate, F1 Score, and Accuracy Score parameters [31] representation. The X-axis of the graph represents different image captioning techniques as discussed in the Resultant Table II. The Y-axis of the graph represents different parameter values. From this graph, it can be concluded that the Precession Rate parameter can give its highest values for GTSCD [6] techniques whereas the Precession Rate parameter gives its lowest value for CDEM [8] method. It can be also concluded that the Recall Rate parameter can give its highest values for GTSCD [6] techniques whereas the Recall Rate parameter gives its lowest value for CDEM [8] method.

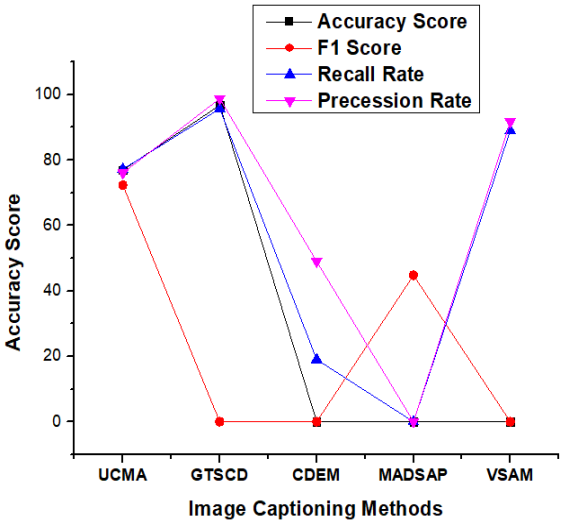

Fig. 9. Shows the Graph-VI

It can be also concluded from the Graph-VI that the F1 Score parameter can give its highest values for UCMA [1] techniques whereas the F1 Score parameter gives its lowest value for MADSAP [14] method. It can be also concluded from the Graph-VI that the Accuracy Score parameter can give its highest values for GTSCD [6] techniques whereas

the Accuracy Score parameter gives its lowest value for UCMA [1] method. The hyphen symbol in the Resultant Table I and II indicates that the said parameter is not computed for the indicated method. Resultant Table II represents some of the methods which are non-computed for the parameters like BLEU, METEOR, ROUGE, CIDER, and SPICE. Resultant-Table-I represents some of the methods which are non-computed for the parameters like Precession Rate, Recall Rate, F1 Score, and Accuracy Score.

Figure 10 and **Figure 11** represent different images along with their captions from the dataset. These images are considered to be the output of the techniques as discussed in the Resultant Table I and Resultant Table II.

In **Figure 10** different methods from the Resultant Table I have displayed, along with the resultant caption, and the method name for which it is generated. In **Figure 10** and **Figure 11** different methods from the Resultant Table I and Resultant Table II are displayed, along with the resultant caption, and the method name for which it is generated.

5. Conclusion

This article analyzes different image captioning techniques implemented over the last decade. First, different methodologies are analyzed concerning their applications, algorithms, merits, and demerits. Then these methods are compared concerning the usage of various datasets, and different parameters to measure their performance. At last, their results in terms of different parameters are analyzed. This article is capable of providing a complete structural analysis of various image captioning techniques in recent years. The progress of image caption generation in the last few decades becomes absolutely clear to any researchers if the whole article is studied properly. In the future, this analytical study will help the researchers to produce a robust technique that may help to overcome all the flaws of the existing techniques and helps to inherit all the features of them.

Author contributions

Name1 Surname1: Conceptualization, Methodology, Software, Field study **Name2 Surname2:** Data curation, Writing-Original draft preparation, Software, Validation., Field study **Name3 Surname3:** Visualization, Investigation, Writing-Reviewing and Editing.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] P. T. Pham, M. Moens, and T. Tuytelaars, "Cross-Media Alignment of Names and Faces," in IEEE Transactions on Multimedia, vol. 12, no. 1, pp. 13-27, Jan. 2010, doi: 10.1109/TMM.2009.2036232.

[2] S. Ye, J. Han, and N. Liu, "Attentive Linear Transformation for Image Captioning," in IEEE Transactions on Image Processing, vol. 27, no. 11, pp. 5514-5524, Nov. 2018, doi: 10.1109/TIP.2018.2855406.

[3] C. C. Park, B. Kim and G. KIM, "Towards Personalized Image Captioning via Multimodal Memory Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 4, pp. 999-1012, 1 April 2019, doi: 10.1109/TPAMI.2018.2824816.

[4] C. Yan et al., "Task-Adaptive Attention for Image Captioning," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 1, pp. 43-51, Jan. 2022, doi: 10.1109/TCSVT.2021.3067449.

[5] A. Karpathy and L. Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, pp. 664-676, 1 April 2017, doi: 10.1109/TPAMI.2016.2598339.

[6] A. Javed, K. B. Bajwa, H. Malik, and A. Irtaza, "An Efficient Framework for Automatic Highlights Generation from Sports Videos," in IEEE Signal Processing Letters, vol. 23, no. 7, pp. 954-958, July 2016, doi: 10.1109/LSP.2016.2573042.

[7] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring Models and Data for Remote Sensing Image Caption Generation," in IEEE Transactions on Geoscience and Remote Sensing, vol. 56, no. 4, pp. 2183-2195, April 2018, doi: 10.1109/TGRS.2017.2776321.

[8] A. Tariq and H. Foroosh, "A Context-Driven Extractive Framework for Generating Realistic Image Descriptions," in IEEE Transactions on Image Processing, vol. 26, no. 2, pp. 619-632, Feb. 2017, doi: 10.1109/TIP.2016.2628585.

[9] N. Yu, X. Hu, B. Song, J. Yang, and J. Zhang, "Topic-Oriented Image Captioning Based on Order-Embedding," in IEEE Transactions on Image Processing, vol. 28, no. 6, pp. 2743-2754, June 2019, doi: 10.1109/TIP.2018.2889922.

[10] N. Xu et al., "Multi-Level Policy and Reward-Based Deep Reinforcement Learning Framework for Image Captioning," in IEEE Transactions on Multimedia, vol. 22, no. 5, pp. 1372-1383, May 2020, doi: 10.1109/TMM.2019.2941820.

[11] Z. -J. Zha, D. Liu, H. Zhang, Y. Zhang and F. Wu, "Context-Aware Visual Policy Network for Fine-Grained Image Captioning," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no.

2, pp. 710-722, 1 Feb. 2022, doi: 10.1109/TPAMI.2019.2909864.

[12] M. Yang et al., "Multitask Learning for Cross-Domain Image Captioning," in IEEE Transactions on Multimedia, vol. 21, no. 4, pp. 1047-1061, April 2019, doi: 10.1109/TMM.2018.2869276.

[13] L. Cheng, W. Wei, X. Mao, Y. Liu, and C. Miao, "Stack-VS: Stacked Visual-Semantic Attention for Image Caption Generation," in IEEE Access, vol. 8, pp. 154953-154965, 2020, doi: 10.1109/ACCESS.2020.3018752.

[14] Y. Huang, J. Chen, W. Ouyang, W. Wan, and Y. Xue, "Image Captioning With End-to-End Attribute Detection and Subsequent Attributes Prediction," in IEEE Transactions on Image Processing, vol. 29, pp. 4013-4026, 2020, doi: 10.1109/TIP.2020.2969330.

[15] B. Wang, C. Wang, Q. Zhang, Y. Su, Y. Wang and Y. Xu, "Cross-Lingual Image Caption Generation Based on Visual Attention Model," in IEEE Access, vol. 8, pp. 104543-104554, 2020, doi: 10.1109/ACCESS.2020.2999568.

[16] J. Ji, C. Xu, X. Zhang, B. Wang and X. Song, "Spatio-Temporal Memory Attention for Image Captioning," in IEEE Transactions on Image Processing, vol. 29, pp. 7615-7628, 2020, doi: 10.1109/TIP.2020.3004729.

[17] J. Wu, T. Chen, H. Wu, Z. Yang, G. Luo, and L. Lin, "Fine-Grained Image Captioning With Global-Local Discriminative Objective," in IEEE Transactions on Multimedia, vol. 23, pp. 2413-2427, 2021, doi: 10.1109/TMM.2020.3011317.

[18] L. Wu, M. Xu, L. Sang, T. Yao, and T. Mei, "Noise Augmented Double-Stream Graph Convolutional Networks for Image Captioning," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 8, pp. 3118-3127, Aug. 2021, doi: 10.1109/TCSVT.2020.3036860.

[19] M. Liu, H. Hu, L. Li, Y. Yu, and W. Guan, "Chinese Image Caption Generation via Visual Attention and Topic Modeling," in IEEE Transactions on Cybernetics, vol. 52, no. 2, pp. 1247-1257, Feb. 2022, doi: 10.1109/TCYB.2020.2997034.

[20] S. Zhang, Y. Zhang, Z. Chen and Z. Li, "VSAM-Based Visual Keyword Generation for Image Caption," in IEEE Access, vol. 9, pp. 27638-27649, 2021, doi: 10.1109/ACCESS.2021.3058425.

[21] D. Hou, Z. Zhao, Y. Liu, F. Chang, and S. Hu, "Automatic Report Generation for Chest X-Ray Images via Adversarial Reinforcement Learning," in IEEE Access, vol. 9, pp. 21236-21250, 2021, doi: 10.1109/ACCESS.2021.3056175.

[22] Z. Zhou et al., "An Image Captioning Model Based on Bidirectional Depth Residuals and its Application," in IEEE Access, vol. 9, pp. 25360-25370, 2021, doi: 10.1109/ACCESS.2021.3057091.

[23] H. Ben et al., "Unpaired Image Captioning With semantic-Constrained Self-Learning," in IEEE Transactions on Multimedia, vol. 24, pp. 904-916, 2022, doi: 10.1109/TMM.2021.3060948.

[24] H. Yanagimoto and M. Shozu, "Multiple Perspective Caption Generation with Attention Mechanism," 2020 9th International Congress on Advanced Applied Informatics (IIAI-AAI), 2020, pp. 110-115, doi: 10.1109/IIAI-AAI50415.2020.00031.

[25] A Guide to Image Captioning, Proteinatlas web resource, https://towardsdatascience.com/a-guide-to-image-captioning-e9fd5517f350, Accessed 05 February 2022.

[26] Papineni, Kishore, et al. "BLEU: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002.

[27] Banerjee, Satanjeev, and Alon Lavie. "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments." Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005.

[28] Lin, Chin-Yew. "ROUGE: A package for automatic evaluation of summaries." Text Summarization Branches Out, 2004.

[29] Anderson, Peter, Basura Fernando, Mark Johnson, and Stephen Gould. “SPICE: Semantic Propositional Image Caption Evaluation.” ECCV (2016).

[30] Vedantam, Ramakrishna, C. Lawrence Zitnick, and Devi Parikh. "CIDER: Consensus-based image description evaluation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[31] Dalianis, H. (2018). Evaluation Metrics and Evaluation. In: Clinical Text Mining. Springer, Cham. https://doi.org/10.1007/978-3-319-78503-5_6.

[32] Cross entropy, Proteinatlas web resource, https://en.wikipedia.org/wiki/Cross_entropy/, Accessed 5th March 2022.

[33] Plausibility, Proteinatlas web resource, https://www.sciencedirect.com/topics/mathematics/plausibility, Accessed 5th March 2022.

[34] What is "Relevance" and how is it calculated?

Proteinatlas web resource, https://dimensions.freshdesk.com/support/solutions/articles/23000022475/, Accessed 5th March 2022.

[35] Tran, Kenneth & He, Xiaodong & Zhang, Lei & Sun, Jian. (2016). Rich Image Captioning in the Wild. 434-441. 10.1109/CVPRW.2016.61.

[36] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3156-3164, doi: 10.1109/CVPR.2015.7298935.

[37] Lin, TY. et al. (2014). Microsoft COCO: Common Objects in Context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8693. Springer, Cham. https://doi.org/10.1007/978-3-319-10602-1_48.

[38] Hsankesara (2018), "Flickr Image dataset", https://www.kaggle.com/datasets/hsankesara/flickr-image-dataset, Accessed on 11th March 2022.

[39] RSICD,https://github.com/201528014227051/RSICD_optimal, Accessed on 11th March 2022.

[40] Lalu Erfandi Maula Yusnu (2021), Oxford 102 Flower Dataset, https://www.kaggle.com/datasets/nunenuh/pytorch-challange-flower-dataset, Accessed on 11th March 2022.

[41] Thomee, Bart & Elizalde, Benjamin & Shamma, David & Ni, Karl & Friedland, Gerald & Poland, Douglas & Borth, Damian & Li, Li-Jia. (2016). YFCC100M: the new data in multimedia research. Communications of the ACM. 59. 64-73. 10.1145/2812802.

[42] Krause, Jonathan & Johnson, Justin & Krishna, Ranjay & Fei-Fei, Li. (2016). A Hierarchical Approach for Generating Descriptive Image Paragraphs.

[43] Raddar (2020), "Chest X-rays (Indiana University)",https://www.kaggle.com/datasets/raddar/chest-xrays-indiana-university, Accessed 11th March 2022.

[44] Johnson, A., Pollard, T., Mark, R., Berkowitz, S., & Horng, S. (2019). MIMIC-CXR Database (version 2.0.0). PhysioNet. https://doi.org/10.13026/C2JT1Q.

[45] "Labeled Faces in the Wild Home", http://vis-www.cs.umass.edu/lfw/, Accessed 11th March 2022.

[46] P Naga Srinivasu, Akash Kumar Bhoi, Rutvij Jhaveri, G Thippa Reddy. Muhammad Bilal, "Probabilistic Deep Q Network for Real-time Path Planning in Censorious Robotic Procedures using Force Sensors", Journal of Real-Time Image Processing, Springer, 2021.

[47] Aditya Khamparia, Deepak Gupta, Victor Hugo C. de Albuquerque, Arun Kumar Sangaiah, Rutvij H. Jhaveri, "Internet of Health Things-driven Deep Learning System for Detection and Classification of Cervical Cells using Transfer Learning", The Journal of Supercomputing, DOI: https://doi.org/10.1007/s11227-020-03159-4 , Springer, Jan 2020.

[48] Jagannath Paramguru, Subrat Kumar Barik, Ajit Kumar Barisal, Gaurav Dhiman, Rutvij H. Jhaveri, Mohammed Alkahtani, Mustufa Haider Abidi, "Addressing Economic Dispatch Problem with Multiple Fuels using Oscillatory Particle Swarm Optimization", Computers, Materials & Continua (CMC, ISSN: 1546-2218), Tech Science Press, Aug 2021.

[49] Surono, S., Rivaldi, M., Dewi, D. A., & Irsalinda, N. (2023). New Approach to Image Segmentation: U-Net Convolutional Network for Multiresolution CT Image Lung Segmentation. Emerging Science Journal, 7(2), 498-506.

[50] Mr. Rahul Sharma. (2013). Modified Golomb-Rice Algorithm for Color Image Compression. International Journal of New Practices in Management and Engineering, 2(01), 17 - 21. Retrieved from http://ijnpme.org/index.php/IJNPME/article/view/13

[51] Arularasan, A. N. ., Aarthi, E. ., Hemanth, S. V. ., Rajkumar, N. ., & Kalaichelvi, T. . (2023). Secure Digital Information Forward Using Highly Developed AES Techniques in Cloud Computing. International Journal on Recent and Innovation Trends in Computing and Communication, 11(4s), 122–128. https://doi.org/10.17762/ijritcc.v11i4s.6315

[52] Ms. Pooja Sahu. (2015). Automatic Speech Recognition in Mobile Customer Care Service. International Journal of New Practices in Management and Engineering, 4(01), 07 - 11. Retrieved from http://ijnpme.org/index.php/IJNPME/article/view/34

[53] Diniesh, V. C. ., Prasad, L. V. R. C. ., Bharathi , R. J. ., Selvarani, A., Theresa, W. G. ., Sumathi, R. ., & Dhanalakshmi, G. . (2023). Performance Evaluation of Energy Efficient Optimized Routing Protocol for WBANs Using PSO Protocol. International Journal on Recent and Innovation Trends in Computing and Communication, 11(4s), 116–121. https://doi.org/10.17762/ijritcc.v11i4s.6314

Table. 1 Comparative Table - I

Year	Author(s)	Main Methodology	Dataset used	Parameter used to calculate performance
2010	Phi The Pham	Unsupervised Cross-Media Alignment	Labeled Faces in TheWild	Recall, Precision, F1-Score
2015	Senmao Ye	Attentive Linear Transformation	MS-COCO, Flickr	BELU-1, BELU-2, BELU-3, BELU-4, METEOR, ROUGE CIDER
2015	Cesc Chunseong Park	Context Sequence Memory Network (CSMN)	InstaPIC-1.1M, YFCC100M	BELU-1, BELU-2, BELU-3, BELU-4, METEOR, ROUGE CIDER, F1 score, Plausibility, Relevance
2015	Chenggang Yan	Task-Adaptive Attention Module	MS-COCO	BELU-1, BELU-2, BELU-3, BELU-4, METEOR, ROUGE CIDER
2015	Andrej Karpathy	Visual-Semantic Alignment	Flickr, and MS-COCO	BELU-1, BELU-2, BELU-3, BELU-4, METEOR, ROUGE CIDER
2016	Ali Javed	Gradual Transition and Scope-Caption Detection	Sports video dataset	Accuracy Rate
2017	Xiaoqiang Lu	Multimodal Method	RSICD	BELU-1, BELU-2, BELU-3, BELU-4, METEOR, ROUGE CIDER
2017	Amara Tariq	Context-Driven Extractive Method	news image dataset	Mean Precision, Mean Recall
2018	Niange Yu	Topic Oriented Model	MS-COCO, Flickr	BELU-1, BELU-2, BELU-3, BELU-4, METEOR, ROUGE CIDER
2018	Ning Xu	Multi-Level Policy and Reward RL Framework	Flickr, MS-COCO	BELU, METEOR, ROUGE CIDER
2019	Zheng-Jun Zha	Context-Aware Visual Policy Network (CAVP)	MS-COCO, Stanford image paragraph captioning dataset	BELU-1, BELU-2, BELU-3, BELU-4, METEOR, ROUGE CIDER
2020	Min Yang	Multitask Learning Algorithm for cross-Domain Image Captioning (MLADIC)	MS-COCO, Flickr, Oxford-102	BELU-1, BELU-2, BELU-3, BELU-4, METEOR, ROUGE CIDER
2020	Yiqing Huang	Multimodal Attribute Detector and Subsequent Attribute Predictor	MS-COCO	BELU-1, BELU-2, BELU-3, BELU-4, METEOR, ROUGE CIDER
2020	Bin Wang	Visual Attention Model	Flickr	BELU-1, BELU-2, BELU-3, BELU-4, METEOR, ROUGE CIDER
2021	Suya Zhang	Visual Semantic Attention Model (VSAM)	IVKD	Precision, Recall
2021`	Daibing Hou	Adversarial Reinforced Report-Generation Framework	IU X-Ray, MIMIC-CXR	BELU-1, BELU-2, BELU-3, BELU-4, METEOR, ROUGE CIDER
2021	Huixia Ben	Semantic-Constrained Self-learning (SCS)	images from Flickr and captions from MS-COCO	BELU-1, BELU-2, BELU-3, BELU-4, METEOR, ROUGE CIDER

Table. 2. Resultant Table I

Method	BELU-1	BELU-2	BELU-3	BELU-4	METEOR	ROUGE	CIDER	SPICE
ALT [2]	75.10	59.00	45.70	35.50	27.40	55.90	110.70	20.30
CSMN [3]	57.10	60.80	42.90	21.30	26.40	57.70	121.40	-
TAAM [4]	79.00	64.10	49.70	37.80	27.50	57.60	116.20	-
VSA [5]	62.50	45.00	32.10	23.00	19.50	-	66.00	-
MULTIMODAL[7]	50.03	31.95	23.19	17.77	20.46	43.33	118.01	-
CDEM [8]	-	-	-	-	25.30	-	-	-
TOM[9]	74.00	56.70	43.40	31.40	25.50	53.40	98.30	_
MLPN [10]	80.20	63.30	48.10	35.70	27.10	57.00	114.10	-
CAVP [11]	80.10	64.70	50.00	37.90	28.10	58.20	121.60	-
MLADIC[12]	79.40	63.10	48.20	36.10	28.10	57.50	119.60	-
STACK-VS [13]	79.00	63.40	48.90	37.20	27.80	57.50	118.90	-
MADASAP [14]	**80.50**	**65.10**	**50.40**	**38.60**	28.70	**58.50**	**128.80**	**22.20**
VAM [15]	73.10	55.90	43.40	32.80	25.49	-	95.10	-
STMA [16]	80.30	64.60	49.80	37.70	28.30	58.10	123.10	-
GLDO [17]	-	-	-	36.30	27.20	27.80	121.10	21.60
NADGCN [18]	77.00	64.00	49.20	36.10	27.80	57.00	116.10	21.30
NICVATP2L [19]	67.10	46.80	34.40	23.50	**30.30**	52.10	66.90	-
ARRGF [21]	-	-	-	14.80	25.30	32.90	40.20	-
BDRGRUN [22]	63.10	44.30	30.10	20.70	20.30	46.20	65.70	13.10
SCS [23]	67.10	47.90	33.40	22.10	20.90	47.30	72.20	13.70

Table. 3. Resultant Table II

Method	Precession Rate	Recall Rate	F1 Score	Accuracy Score
UCMA [1]	76.12	77.21	**72.33**	77.00
GTSCD [6]	**98.80**	**95.70**	**-**	**96.78**
CDEM [8]	49.00	19.00	-	-
MADSAP [14]	-	-	44.80	-
VSAM [20]	91.70	89.02	-	-

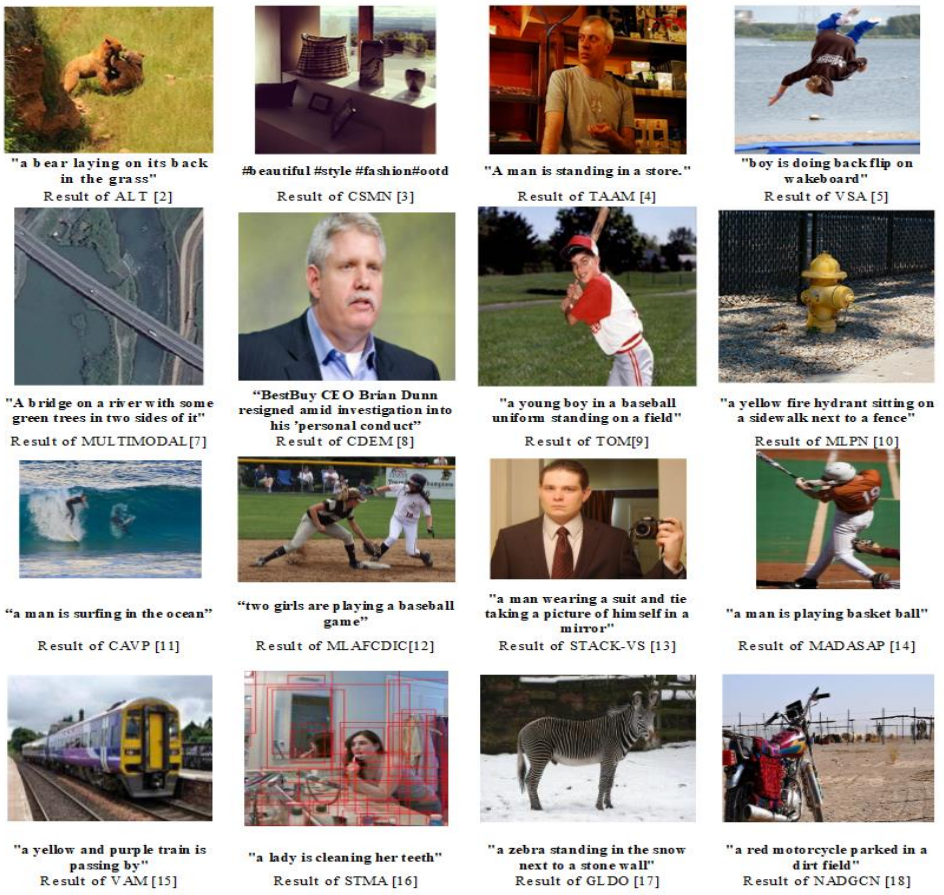

Fig. 10. Shows the generated captions from Resultant Table I

Fig. 11. Shows the generated captions from Resultant Table I and Resultant Table II