Survey on Image Description Generation using Deep Learning

Jay Pachupate
Student
Wakad
Pune - 411057

Sonal Fatangare
Guide
Warje
Pune - 411058

Sneha Tambare
Student
Katraj
Pune - 411046

Shraddha Debadwar
Student
Kondhwe Dhawde
Pune - 411023

Pratik Sonar
Student
Kothrud
Pune – 411038

ABSTRACT

In the future, the creation of an image description system could aid those who are blind or visually handicapped in "perceiving" the world. In natural language processing and computer vision, producing logical and contextually appropriate written descriptions for images is a crucial task that is referred to as "image description generation". In Bi-LSTM processes input sequences sequentially and captures contextual information in a bidirectional manner, but it may not capture long-range dependencies effectively. To tackle these issues, the survey focuses on leveraging the BERT uses a self-attention mechanism that allows it to capture context over much longer ranges, making it more effective in handling global context and dependencies. For the model to comprehend the visual elements, BERT must be integrate contextual relationships between different visual elements and generate more coherent and contextually relevant image descriptions.

Keywords

Convolutional Neural Network (CNN), Bidirectional Encoder Representations from Transformers (BERT), Fine-tuning, Evaluation Metrics, and Benchmark Datasets.

1. INTRODUCTION

In recent years, computer vision research has shifted its focus to more complex tasks, especially in the area of generating captions for images. While people can easily describe the content of images, teaching machines to do the same has been a tough challenge. The research direction has evolved from successes in visual classification to exploring neural machine translation models for image description generation. These models, originally designed for language translation, have shown great success in helping machines understand the meaning of images and generate coherent and contextually relevant descriptions.

Many translation models employ an encoder-decoder architecture, in which a fixed-length vector is created by the encoder compressing information from the source language. The decoder then makes use of this vector to produce a phrase in the language of choice. This approach works well when the structure and phrases of source and target sentences are similar. However, when applied to generating captions for images, which have very different structures compared to sentences, challenges arise. In image caption generation, the algorithm needs to compress detailed visual information into a vector during the encoding phase. This raises concerns about the model's ability to capture the nuanced structural details needed for generating accurate and contextually rich textual descriptions during the decoding phase.

This survey study undertakes a thorough investigation of the use of the BERT algorithm in the field of picture description creation. The BERT algorithm is used in generating descriptions for images. BERT is well-known for its ability to understand words in context and pay attention to both sides of a sentence. It seems promising for addressing the differences between how we describe things in words and how we show them in pictures. With the goal of providing a thorough analysis of BERT's strengths and potential weaknesses, this survey delves deeply into the use of BERT to generate captions for pictures. By looking at existing research and real-world results, the survey aims to share useful insights into how BERT is currently used for describing images and where it might go in the future.

2. LITERATURE SURVEY

A literature survey is a critical analysis of existing research. It provides a framework for comprehending the level of knowledge at the moment in a certain topic and aids in locating holes, employed algorithms, and potential research fields. A literature survey involves a systematic examination of a body of published works. The primary goal is to provide a comprehensive overview of Published papers. It involves dataset used for the work, feature extraction, Classification and evaluation metrics. It also summarizes the research gap for research.

In 2023 [1] HUAWEI ZHANG et al developed Contextual Data combination Using Bi-LSTM for Image Caption Generation. The proposed method uses a Bi-LSTM to anticipate visual content based on context cues. This structure saves future information in addition to past information.The forward decoder and the backward decoder receive the visual data respectively, and from them they extract semantic information and produce corresponding semantic output.Using the MS-COCO dataset, they assessed the suggested approach and gave it a BLEU-4 score of 37.5.

Qimin Cheng et al [3] developed the Sydney-Captions and UCM-Captions benchmark datasets, as well as a deep multimodal neural network model. To increase performance, their model combined RNN/LSTMs with several CNNs.In order to provide more accurate and adaptive descriptions, this study built a bigger benchmark dataset, RSICD, and analyzed and summarized the difficulties associated with captioning remote sensing images.

In 2021 [5] KYUNGBOK MIN et al created Encoder-Decoder Architecture based Deep Learning Short Tale Generation for an Image. This work uses a common image caption dataset and a manually acquired story corpus to build a SSCap. They used the popular MS-COCO dataset to assess the proposed method, obtaining a CIDEr Score of 53%.In this study, the CNN and RNN algorithms are used.

The resulting model in 2021 [7] regulated both imageability and duration, allowing output to be customized for different uses. Experiments demonstrated that the system could regulate the length and visual descriptiveness of the generated captions while maintaining a captioning performance comparable to comparison approaches. The goal imageability significantly correlated with human expectations, according to a subjective evaluation conducted with human participants. As a result, the system verified that the suggested approach offered a positive first step towards customizing picture captions for specific applications. The BLEU Score is 56.075.

In 2020 [10] suggested a CNN-based deep learning method for captioning images in Hindi. After converting an image to a compact representation, it used RNN and LSTM to create an analogous Hindi text. A neural network model with encoder-decoder functionality that can automatically view a picture and Using the Flickr 8k-Hindi datasets, the system was trained to generate a passable description in basic Hindi. The research' findings showed that in contrast to a model the model was trained using five uncleaned descriptions for each image. A caption with only one clear description for each image is of higher quality. Since the model gets the visual characteristic and the word before it, it is trained to predict words one by one. With a BLEU-1 score of 0.585, they assessed the suggested approach using the popular Flickr 8k-Hindi dataset. This study makes use of the CNN and RNN-LSTM algorithms.

In 2020 [11] Jie Wu, et al developed a unique discriminative global-local target to help with the creation of fine-grained descriptive captions.More specifically, they developed a novel global discriminative constraint that, when used worldwide, pulls the generated sentence to improve the dataset's capacity to discriminate between the matched image and every other image.To enhance attention to the less common but more specific words/phrases, a local discriminative constraint is suggested from the local perspective. This will make it simpler to write captions that better convey the subtleties of the images that are supplied visually. Using the popular MS-COCO dataset, they assessed the suggested technique and obtained a CIDEr Score of 121.1.

Chinese Event Identification Using BiLSTM and Multi-Feature Merge :Their event extraction technique, which they developed,is becoming more and more crucial for automatically extracting relevant information from vast amounts of unstructured texts. The method divides the context of a word into groups at the sentence and document levels. The contextual data is captured using the BiLSTM model. Simultaneously, a word representation technique appropriate for classification tasks including trigger words is suggested. The BiLSTM model receives word vectors from the phrase one after the other until output vectors with phrase-level context data are produced. It was trained on CEC corpus datasets with CIDEr score 98%, as demonstrated by the experimental results[14].

Peter Anderson, el al [15] created from the bottom-up and top to bottom Attention for Visual Problem Responding and Image Captioning to allow for more in-depth study of images and even more reasoning stages. This paper suggests utilizing this mechanism to calculate attention at the object and other salient regions of the image levels. In this method, it establishes feature weightings and suggests image area using a bottom-up approach, each of which has a matching feature vector. Using this method for captioning images, With MSCOCO test, model achieved CIDEr, SPICE, BLEU-4 grades of 117.9, 21.5, and 36.9, respectively.

Linghui Li, et al[18] since the majority of methods in use today solely use image-level representation, there have been two issues that need to be resolved in recent years. One issue was that it failed to identify certain significant objects while creating the image description, and the other was that it misjudged whether to classify one thing under the incorrect heading. In this suggested work, they presented a novel technique for producing visual descriptions termed global-local attention (GLA) to overcome the two issues. The proposed GLA model used an attention mechanism to integrate characteristics at the object and image levels. In this way, the model was able to simultaneously pay attention to context information and objects with preference. As a result, the suggested GLA technique produced more pertinent picture description sentences and attained results using a number of widely used assessment metrics, including CIDEr, METEOR, ROUGE-L, and BLEU.The CIDEr Score is 96.4%.

A newly created SCA-CNN which combines channel-wise and spatial attention in a CNN is proposed.When captioning images, SCA-CNN continuously modifies the multi-layer feature map's sentence generating context to specify what and where the subject of visual attention. Examine the proposed SCA-CNN architecture with three benchmark datasets of captioned images: Flickr8K, Flickr30K, and MSCOCO. SCA-CNN regularly outperforms the most advanced visual attention-based picture captioning methods significantly [19].

In 2017 [20] Ting Yao, et al developed Boosting Image Captioning with Attributes for using natural language to automatically describe an image. There are five techniques for picking attributes during preprocessing. 1. LSTM-A1 (Only Utilizing Attributes) 2.LSTM-A2 (put the image here first) 3.LSTM-A3 (primary feeding qualities) 4.LSTM-A4 (Image Input at Every Time Step) 5.LSTM-A5 (Attributes input at every time step). LSTM. Using the MS-COCO dataset, they obtained a CIDEr score of 100.2%.In this study CNN, RNN, and LSTM algorithms are applied.

Pranay Mathur, el al [21] created a system of encoders and decoders, with the encoder RNN with LSTM Cells and the decoder a pretrained InceptionV4 Convolutional Neural Network from Google. They demonstrated the real-time applicability and optimisations by implementing an Android application and by using TensorFlow framework.With a CIDEr score of 74.7, the model was developed using the Flickr-30k and MS-COCO dataset.

Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning developed an encoder-decoder system that gave the decoder a fallback option.They also unveiled a brand-new LSTM extension, creating a second "visual sentinel".They experiment method on the MS-COCO image captioning 2015 challenge and Flickr-30k dataset. The model performed noticeably better, increasing the CIDEr score on Flickr30k and COCO from 0.493/1.029 to 0.531/1.085, respectively[22].

In 2017 [23] the problem of using reinforcement learning to enhance image caption systems, and show that by appropriately optimizing systems using the test metrics of the MS-COCO task, model could potentially achieve significant performance gains. With an improvement in the CIDEr score from 104.9 to 114.7, the MSCOCO evaluation established a new benchmark for the work.

In their proposed work, they presented a new technique called semantic attention. The program improved the ability of recurrent neural networks to compile and focus on certain semantic concept ideas into hidden states and outputs. Selection and fusion work together to produce a feedback loop, that combines top-down and bottom-up computation. MS-COCO and Flickr30k, two publicly accessible benchmarks, were used to test their methodology. The algorithm consistently outperformed across a wide variety of evaluation metrics, according to experimental results. The computed CIDEr score is 93.5% [24].

In 2015 [27] Xu Jia, et al LSTM model was extended in this work is referred to as gLSTM for short .To direct the model toward solutions that are more directly related to the image content, the researchers specifically included extra input of semantic information obtained from the image to each unit of the LSTM block. For beam search, different length normalization techniques were also explored to prevent bias around short sentences. Work on several benchmark datasets, such as Flickr8K, Flickr30K, and MS-COCO. The CIDEr score is 81.26.

In 2015 [31] suggested a neural-image convolution network that reduced a picture to a little representation and then used a RNN to generate a phrase that matched. The model was trained with the image in order to optimize the chance of the text. Tests conducted on multiple datasets, including the Pascal dataset, Flickr30k, and MS-COCO, demonstrated the resilience of NIC with a CIDEr score of 85.5.

Table - 1: Deep Literature Survey of Current Technologies

Ref. No.	Year	Data Set	Feature Extraction and Classification	Quality metrics	Research Gap Identified
[1]	2023	MS-COCO dataset	CNN, BI-LSTM(F-LSTM & B-LSTM)	BLEU-4 = 37.5	Global constraints for detailed semantics
[2]	2023	Flickr30, MSCOCO	optical character recognition, RNN	2.7 points on the CIDEr score	Enhancing English OCR
[3]	2022	UCM-Captions,Sydney-Captions,RCISD,NWPU-Captions	CNN, RNN/LSTM	Range of BLEU, METEO, ROUGE, and SPICE falls between 0 and 1, and the range of CIDEr is 0 and 5.	Cross-modal retrieval for multisource data.
[4]	2022	Stanford image paragraph	MEG and PaG(Encoder and decoder), MEG and PaG(Encoder and decoder)	CIDEr = 63%	Could integrate a sophisticated ML model
[5]	2021	MS-COCO	CNN, RNN	CIDEr = 53%	Could look other Transformer-based architectures
[6]	2021	Microsoft COCO and Flickr30k	Wordnet, Imageability Dictionary.	TAYI = 36%	Could improve the second step of the SD-RSIC
[7]	2021	UIUC phrasal recognition, six-class sport, Microsoft COCO.	Faster RCNN, FTWIM	BLEU = 56.075	Format and optimization long and complex sentences
[8]	2021	UCM Captions, RSICD Data sets, Sydney Caption	CNN, RNN, ImageNet	26% for the RSICD data set in CIDEr	Could Enhance the SD-RSIC's Phase 2
[9]	2020	AIC-ICC image dataset	attention mechanism, two-layer long short-term memory (LSTM)	CIDEr =98%	Could directly use the evaluation metric as an optimization objective function
[10]	2020	Flickr 8k-Hindi	CNN, RNN-LSTM	BLEU-1 score of 0.585	Could study on wide-Ranging applications RFID technology
[11]	2020	MS-COCO	CNN with (TDA +GLD),LSTM	CIDEr = 121.1	Format and optimization long and complex sentences

[12]	2020	MSCOO, Flickr 8k/Flickr30k, PASCAL, AIC AI Challenger and STAIR .	SCA-CNN,RNN	CIDEr = 88.5	Could study on An image that is often rich in content.
[13]	2019	Microsoft COCO and Flickr30k	CNN, Dynamic greatest pooling BiLSTM-CNN architecture based on portions.	CIDEr = 57.69%	Could establish inter-dependencies between higher hierarchical discourse units
[14]	2019	CEC corpus	part of speech tagging. ,BI- LSTM	F1 score = 0.778	Could improve localizati0n accuracy
[15]	2018	MSCOCO, VQA v2.0	CNN, R-CNN	CIDEr = 117.9%	Could use language model to refine and polish the generated captions
[16]	2018	PASCAL VOC and MS COCO	CNN, ResNet	mAP score=78.6	Could address small object detection
[17]	2018	MSCOCO, Flickr 30K, MNIST, PASCAL VOC, DAQU R	CNN, DNN, GRU/RNN, LSTM	CIDEr = 109.1	Could improve caption diversity
[18]	2017	MS COCO	Deep CNN, R-CNN method, Combined local and global aspects through the use of an attention mechanism.	CIDEr = 96.4%	Possibly combine the image feature extractor and object detector
[19]	2017	Flickr 8k, Flickr30, MSCOCO	CNN, RNN/LSTM	SCA-CNN only made up 0.1% in METEOR and 0.6% in BLEU4.For the MSCOCO server result	Must look for ways to add more watchful layers without going overboard.
[20]	2017	MSCOCO	CNN, RNN/LSTM	CIDEr = 100.2%	Could refer to the concept of reinforcement learning algorithm
[21]	2017	Flickr30k, MS-COCO	Inception V4 CNN by Google, Deep RNN with LSTM	CIDEr = 74.7	could explore for user-specific customization options
[22]	2017	The MS-COCO image captioning 2015 challenge dataset and Flickr30K.	CNN, RNN with LSTM along with visual sentinel	CIDEr = 0.531 for Flickr30k CIDEr =1.085for MS-COCO	Look for advancement and effectiveness of image captioning models.
[23]	2017	MSCOCO	CNN, RNN/LSTM	CIDEr from 104.9 to 114.7	Improvements in captioning performance
[24]	2016	Microsoft COCO and Flickr30K.	semantic attention, RNN	CIDEr=93.5%	Could explore new models for proposed semantic attention mechanism.
[25]	2016	PASCAL, MS- COCO IMAGE NET	Residual networks, Image-Net	CIDEr=133.3%	Could combine with stronger regularization to improve results.

[26]	2016	FlickrLogos-32	MSER algorithm,Bag-of-words (BOW) and GoogLeNet	recall=64.7%, precision=4.7% and f-score=8.7%	Could enhance performance
[27]	2015	Flickr 8K, Flickr30K and MS COCO	Semantic data that was taken out of the picture and added to each LSTM block unit.	CIDEr for gLSTM Min hinge=81.26	Could further improve on performance to be obtained by integrating the schemes.
[28]	2015	Conceptual Captions	Image processing repository tool.	80%	The study provides a fresh perspective for future scholars working in other fields.
[29]	2015	PASCAL-50S, ABSTRACT-50S, MSCOCO	NA	NA	Could provide a benchmark for future comparisons.
[30]	2015	Flickr 8K, Flickr30K, MS-COCO	CNN, Bi-RNN/Multimodal RNN	CIDEr = 66%	Encompass ambiguity handling
[31]	2015	Pascal dataset,Flickr30k,MS-COCO	CNN, RNN	CIDEr = 85.5	Could elevate the overall quality and informativeness of generated image captions.
[32]	2013	Flickr Dataset	ILP	BLEU = 0.29%	Extend technique to manage broader image
[33]	2012	ImageNet	Feature Extraction using image parsing, Classification using SVM	CIDEr = 63%	Could be migrated to other content recognitions in image

3. ALGORITHMIC SURVEY

An algorithm survey is a comprehensive examination and analysis of existing algorithms within a specific domain or field. It involves the systematic review of various algorithms, features, performance metrics, advantages, disadvantages and other relevant information for each algorithm, offering readers an overview of the cutting edge in this domain. The temporal complexity is expressed as O (N.M+L.P), in which N represents dimensions of input images, M denotes factors related to image feature extraction (e.g., CNN complexity), L relates to language model complexity (e.g., RNN or transformer), and P includes other processing considerations. Space complexity would be influenced by factors like model size, input size, and intermediate activations.

Table - 2: Algorithmic Survey of Research Studies

Ref. No.	Year	Algorithm Used	Evaluation Metrics
[1]	2023	CNN & BI-LSTM(F-LSTM & B-LSTM)	BLEU-4 = 37.5
[2]	2023	Optical character recognition, RNN	2.7 points on the CIDEr score
[3]	2022	CNN, RNN, LSTM	Diversity of BLEU, METEOR, ROUGE, and SPICE is between 0 and 1, and the range of CIDEr is 0 and 5.
[4]	2022	MEG and PaG(Encoder and decoder)	CIDEr = 63%
[5]	2021	CNN, RNN, Adam Algorithm	CIDEr = 53%
[6]	2021	CNN, BERT	TAYI = [36]
[7]	2021	Faster RCNN and FTWIM	BLEU* = 56.075
[8]	2021	CNN, RNN, LSTM, SD-RSIC	CIDEr = 26%
[9]	2020	CNN, LSTM	CIDr = 98%
[10]	2020	CNN, RNN-LSTM	BLEU-1 score of 0.585,
[11]	2020	CNN and LSTM	CIDEr = 121.1
[12]	2020	SCA-CNN, RNN	CIDEr = 88.5
[13]	2019	BI-LSTM, CNN	CIDEr = 57.69%
[14]	2019	BI-LSTM	F1 score = 0.778
[15]	2018	CNN, R-CNN	CIDEr = 117.9%
[16]	2018	CNN, ResNet	mAP score =78.6
[17]	2018	CNN, DNN, GRU/RNN, LSTM	CIDEr = 109.1

[18]	2017	GLA model	CIDEr = 96.4%
[19]	2017	CNN, SCA-NN, LSTM	SCA-CNN only 0.6% in BLEU4 and 0.1% in METEOR, respectively. For the MSCOCO server results, Google NIC surpass Only 0.7% of BLEU4 had SCA-CNN. and 1% in METEOR
[20]	2017	CNN, RNN, LSTM	CIDEr = 100.2%
[21]	2017	Inception V4 CNN by Google, Deep RNN with LSTM	CIDEr = 74.7
[22]	2017	CNN, RNN with LSTM along with visual sentinel	CIDEr = 0.531 for Flickr30k CIDEr = 1.085 for MS-COCO
[23]	2017	CNN, RNN, LSTM, REINFORCE algorithms	CIDEr from 104.9 to 114.7
[24]	2016	Semantic attention algorithm	CIDR = 93.5%
[25]	2016	Residual learning algorithm, Image-Net	CIDEr = 82.3%
[26]	2016	MSER algorithm	Recall = 64.7%, precision = 4.7% and f-score = 8.7%
[27]	2015	gLSTM	CIDEr for gLSTM Min hinge = 81.26
[28]	2015	Prepared image processing repository tool	CIDEr = 53%
[29]	2015	NA	NA
[30]	2015	CNN,Bi-LSTM, Multimodal LSTM	CIDEr = 66%
[31]	2015	CNN and RNN	CIDEr = 85.5
[32]	2013	SVM, RFM	BLEU = 0.29%
[33]	2012	Graph based algorithm and SVM	CIDEr = 63%

3.1 Text Generation

Text generation is the process by which an AI system creates written content by imitating human linguistic patterns and styles. The procedure entails writing relevant, cogent language that imitates informal human communication. In a number of fields, including content creation, code support, natural language processing, and customer service, text generation has grown in significance.

3.1.1 Quality Metrics

The effectiveness and performance of a statistical or machine learning model are assessed using quality metrics, which are numerical measurements.These metrics provide information about the model's performance and facilitate the comparison of other models or algorithms. Various quality metrics are used in the machine learning model such as CIDEr, BLEU, ROUGE, METEOR, etc.

3.1.2 Dataset

A dataset is an assemblage of several kinds of digitally stored data that is utilized for text preparation Various datasets are used in model like MSCOCO, Flickr 8k, Flickr30k, Pascal, ImageNet.

3.1.3 Deep Learning

Deep learning is an Artificial Intelligence technique that trains computers to use a method of information processing equivalent to the human brain. Deep learning algorithms can recognize complex patterns in photos, text, sounds, and other kinds of data to produce accurate insights and projections. Various Deep learning approaches are CNN, LSTM, RNN, BI-LSTM, BERT etc.

3.1.4 Applications

Application of Text Generation is Image captioning, Text summarization, AMR, Script Generation, Language Generation, Machine Translation, Speech-to-Text, Shopping Guide, Weather Forecast.

3.1.5 Languages

Languages used for Text Generation are English, Chinese, Hindi, Spanish, Russian, Korean.

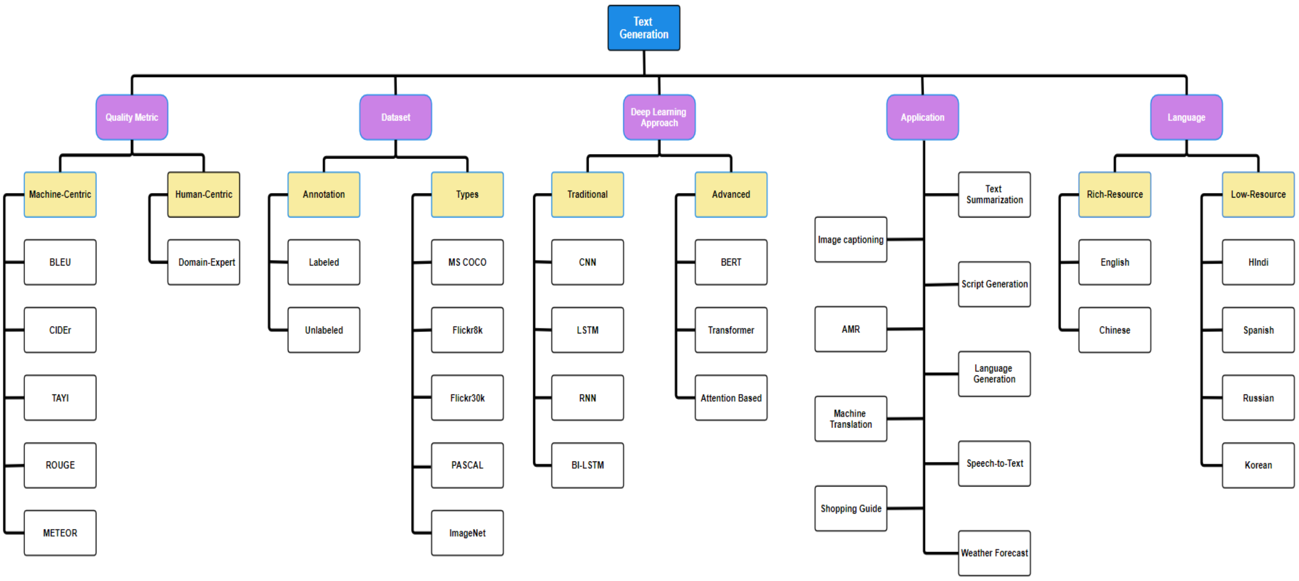

Figure 1. Systematics of the text generation.

4. EQUATIONS

Evaluation metrics are numerical measurements that are used to evaluate the efficacy and performance of machine learning algorithms. These indicators facilitate model comparison and offer insights into the model's performance. Evaluating the performance of image description generation models, especially those incorporating BERT, involves using various metrics to assess their effectiveness. Here are some common quality metrics for text generation**:** CIDEr, BLEU ROUGE, METEOR.

4.1 CIDEr (Consensus-based Image Description Evaluation):

CIDEr, is a statistic used to assess how well automatic image captioning systems produce image descriptions. The goal of CIDEr is to assess how closely the generated captions match reference captions that have been provided by humans. The process of creating natural language descriptions for photos is known as image captioning.

Benchmark datasets for picture captioning, like the MS-COCO dataset, are frequently used along with CIDEr.

Calculation method in CIDEr is given below:

$$\text{CIDEr}_n(c_i,S_i) = \frac{1}{m}\sum j \frac{gn(c_i).gn(s_{ij})}{||gn(c_i)||.||\,gn(s_{ij})||} \qquad (1)$$

$$\text{CIDEr}(c_i,S_i) = \sum_{n=1}^{N} w_n \text{CIDEr}_n(c_i,S_i). \qquad (2)$$

Where $gn(c_i)$ is TF-IDF vector (n-gram),S_{ij} is j th reference,c_i is candidate sentence and S_i is reference sentences.

4.2 BLEU (Bilingual Evaluation Understudy):

BLEU is used to evaluate the level of writing produced by computers, especially when it comes to machine translation.The basic idea behind BLEU is the comparison of n-grams, or contiguous sequences of n elements, typically words, between the machine-generated output and the reference translations.

Calculation method in BLEU is given below:

$$\text{Unigram Precision P}=\frac{m}{w_r} \qquad (1)$$

$$\text{Brevity Precision P} =\begin{cases} 1 & if\ c > r \\ e^{(1-(\frac{r}{c}))} & if\ c <= r \end{cases} \qquad (2)$$

$$\text{BLEU} = \text{p.e} \sum_{n=1}^{N}(\frac{i}{N} * \log P_n) \qquad (3)$$

Here, m is the number of words from the candidate that are included in the reference. Wt is the total number of words in the candidate. c is the translation corpus's total length, and r represents the reference corpus's effective length. The modified n-gram precision's geometric average is denoted by Pn. Pn is calculated using a length of n-grams, N.

4.3 ROUGE (Recall-Oriented Understudy for Gisting Evaluation):

ROUGE is utilized for the automatic assessment of text creation by machines, especially in the context of document and text summarization. ROUGE calculates the amount of overlap that is present between the generated summary and one or more manually-written reference summaries.

In the field of NLP and summarization work, ROUGE is widely used to assess the effectiveness of automated summary systems.

4.4 METEOR (Metric for Evaluation of Translation with Explicit ORdering):

METEOR is an automatic metric that is used, especially in the context of automatic summarization and machine translation, to assess text produced by machines. METEOR has been applied to machine translation, text summarization, and other text production tasks in the background of NLP.

METEOR is Calculated as following where Unigram precision *P* is calculated as:

$$\text{P} = \frac{m}{w_r} \qquad (1)$$

Where is the number of unigrams in the candidate translation and m is the number of unigrams in the candidate translation that also appear in the reference translation. The method for determining the value of Unigram recall R is as follows:

$$R = \frac{m}{w_r} \qquad (2)$$

where is the number of unigrams in the reference translation and m is as stated above. The harmonic mean is used to combine recall and precision in the following way, with recall being weighted nine times higher than precision:

$$F_{mean} = \frac{10PR}{R+9p} \qquad (3)$$

Unigrams are bundled into as few chunks as feasible to calculate this penalty; a chunk is defined as a collection of unigrams that are contiguous in both the hypothesis and the reference. There are fewer chunks the longer the adjacent mappings are en the candidate and the reference. A translation that is exactly the same as the source will yield a single chunk. This is how the penalty p is calculated:

$$P = 0.5(\frac{c}{u_m})^3 \qquad (4)$$

where is the number of unigrams that have been mapped and c is the number of chunks. A segment's ultimate score is determined as M below. In the event that there are no longer or bigger matches, the penalty has the effect of lowering the by up to 50%.

$$M = F_{mean}\ (1 - P) \qquad (5)$$

The aggregate values for P, R, and p are obtained and then merged using the same formula to determine a score for the entire corpus, or collection of segments. The technique can also be used to compare a translation candidate with many translation references. In this instance, the algorithm chooses the candidate with the highest score after comparing it to each of the references.

5. CONCLUSION

In conclusion, a significant improvement is the inclusiveness of the BERT method in image description generating process. Bidirectional attention mechanisms and contextualized embeddings of BERT have been crucial in bridging the structural gap between textual and visual material. Although the study demonstrates that BERT can provide logical and contextually appropriate image captions, further development of its application and investigation of multimodal techniques point to a dynamic future for improving the collaboration between natural language processing and machine vision. While the study shows that BERT can create logical and pertinent image description for the context, more research into multimodal techniques and the application of the technology will likely lead to an increasingly dynamic future for enhancing machine vision and natural language processing collaboration.

Future exploration on BERT based image description generation will provide opportunities to improve fine-tuning techniques for image-centric problems. Further studies may look into integrating BERT with multimodal learning and adding audio and spatial data for captions that are more detailed. In addition, research needs to concentrate on improving interpretability and ethical issues in order to assure the objective and responsible application of image description systems in different situations.

6. REFERENCES

[1] HUAWEI ZHANG, CHENGBO MA, ZHANJUN JIANG, AND JING LIAN ”Image Caption Generation Using Contextual Information Fusion With Bi-LSTM-s.”

[2] ARISA UEDA, WEI YANG, AND KOMEI SUGIURA "Switching Text-Based Image Encoders for Captioning Images With Text"

[3] Qimin Cheng, Haiyan Huang, Yuan Xu, Yuzhuo Zhou, Huanying Li, and Zhongyuan Wang "NWPU-Captions Dataset and MLCA-Net for Remote Sensing Image Captioning"

[4] Thanh-Son Nguyen and Basura Fernando "Effective Multimodal Encoding for Image Paragraph Captioning"

[5] KYUNGBOK MIN, MINH DANG, AND HYUN JOON MOON"Deep Learning-Based Short Story Generation for an Image Using the Encoder-Decoder Structure"

[6] MARC A. KASTNER, KAZUKI UMEMURA, ICHIRO IDE, YASUTOMO KAWANISHI, TAKATSUGU HIRAYAMA, KEISUKE DOMAN, DAISUKE DEGUCHI, HIROSHI MURASE AND SHIN'ICHI SATOH" Imageability- and Length-Controllable Image Captioning"

[7] Lin Huo, Lin Bai, and Shang-Ming Zhou "Automatically Generating Natural Language Descriptions of Images by a Deep Hierarchical Framework"

[8] Gencer Sumbul, Sonali Nayak, and Begüm Demir "SD-RSIC: Summarization-Driven Deep Remote Sensing Image Captioning"

[9] Maofu Liu, Huijun Hu, Lingjun Li, Yan Yu, and Weili Guan "Chinese Image Caption Generation via Visual Attention and Topic Modeling"

[10] Ankit Rathi "Deep learning approach for image captioning in Hindi language"

[11] Jie Wu, Tianshui Chen, Hefeng Wu, Zhi Yang, Guangchun Luo, and Liang Lin "Fine-Grained Image Captioning with Global-Local Discriminative Objective"

[12] Haoran Wang, Yue Zhang, and Xiaosheng Yu "An Overview of Image Caption Generation Methods"

[13] FENGYU GUO, RUIFANG HE, AND JIANWU DANG "Implicit Discourse Relation Recognition via a BiLSTM-CNN Architecture With Dynamic Chunk-Based Max Pooling"

[14] GUIXIAN XU, YUETING MENG, XIAOKAI ZHOU, ZIHENG YU, XU WU, AND LIJUN ZHANG "Chinese Event Detection Based on Multi-Feature Fusion and BiLSTM"

[15] Peter Anderson, Xiaodong He,,Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, Lei Zhang "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering"

[16] Junwei Han, Dingwen Zhang,Gong Cheng, Nian Liu, and Dong Xu "Advanced Deep-Learning Techniques for Salient and Category-Specific Object Detection"

[17] Sidra shabir, Syed Yasser Arafat "An image conveys a message: A brief survey on image description generation"

[18] Linghui Li, Sheng Tang, Member, IEEE, Yongdong Zhang, "GLA: Global-local Attention for Image Description"

[19] Long Chen Hanwang Zhang Jun Xiao Liqiang Nie Jian Shao Wei Liu Tat-Seng Chua"SCA-CNN: Spatial and Channel-wise Attention in Convolutional Networks for Image Captioning"

[20] Ting Yao †, Yingwei Pan ‡, Yehao Li §, Zhaofan Qiu ‡, and Tao Mei † "Boosting Image Captioning with

Attributes”

[21] Pranay Mathur∗, Aman Gill†, Aayush Yadav‡, Anurag Mishra§ and Nand Kumar Bansode “Camera2Caption : A Real-Time Image Caption Generator”

[22] Jiasen Lu, Caiming Xiong, Devi Parikh, Richard Socher “Adaptive Attention via A Visual Sentinel for Image Captioning”

[23] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross and Vaibhava Goel “Self-critical Sequence Training for Image Captioning”

[24] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo “Image Captioning with Semantic Attention”

[25] Kaiming He, Xiangyu Zhang Shaoqing Ren, Jian Sun “Deep Residual Learning for Image Recognition”

[26] Sezer Karaoglu, Ran Taoy, Theo Gevers and Arnold W. M. Smeulders ``Words Matter: Scene Text for Image Classification and Retrieval”

[27] Xu Jia, Efstratios Gavves, Basura Fernando, Tinne Tuytelaars “Guiding the Long-Short Term Memory model for Image Caption Generation”

[28] Priyanka Jain, Priyanka Pawar, Gaurav Koriya, Anuradha Lele, Ajai Kumar and Hemant Darbari “Knowledge acquisition for Language description from Scene understanding”

[29] Ramakrishna Vedantam,C. Lawrence Zitnick,Devi Parikh “CIDEr: Consensus-based Image Description Evaluation”

[30] Andrej Karpathy,Li Fei-Fei “Deep Visual-Semantic Alignments for Generating Image Descriptions”

[31] Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan “Show and Tell: A Neural Image Caption Generator”

[32] Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, Tamara L. Berg “BabyTalk: Understanding and Generating Simple Image Descriptions”

[33] Yan Zhu,Hui Xiang, Wenjuan Feng “Generating Text Description from Content-based Annotated Image”.