Kalpa Publications in Computing

Volume 19, 2024, Pages 303–321

Proceedings of 6th International Conference
on Smart Systems and Inventive Technology

Image Captioning: A Survey on its methods and Implementation.

Dr. R. R. Thirrunavukkarasu[1], Arun A. T[2], Aravind. JG[3], Dharanash. S[4], Kishore. G[5]

[1, 2, 3, 4, 5]Electronics and Communication Engineering, Sri Krishna College Of Technology, Coimbatore, India.
[1]thirrunavukkarasu.r.r@skct.edu.in, [2]20tuec009@skct.edu.in, [3]20tuec007@skct.edu.in, [4]20tuec018@skct.edu.in, [5]20tuec042@skct.edu.in

Abstract

This literature review aims to navigate the vast landscape of image captioning, an interdisciplinary field that lies combining natural language processing and computer vision. We start with a detailed examination of the CNN-to-Bi-CARU model, an attention-based bidirectional architecture for comprehensive contextual information extraction. The application of this model in image captioning therefore necessitates detecting image features and objects, and identifying them precisely. Attention mechanisms are important for securing precise matching regarding changes in focused content during caption generation. The efficiency concerns have been highlighted by the CNN-to-Bi-CARU model that has taken less time in coming up with images during inference. Stability is acknowledged even as improvements are proposed for a perfect BDR-GRU system. The experimental phase investigates different loss functions and optimizers leading to selecting cross-entropy as a loss function and Adam optimizer to achieve BLEU-4 metrics and better accuracy. The introduction of a new framework allows for the estimation of significant regions in images. The approach relies on image captioning, which incorporates semantic information while estimating important regions on basis of subject and object words contained in those captions. Experimental results confirm that the technique can estimate important regions with sensitivity rivaling human perception. In regard to remote sensing image captioning, this exploration ends up with an encoder-decoder model. Instead of traditional token generation, the model supports continuous output representations, using a proposed loss function to optimize semantic similarity at sequence level. This novel way may have a great impact on language generation in the context of remote sensing imagery. Viewing the diverse methods that were explored, problems that have been identified and inventions that have been realized, this paper provides an overview of the landscape and a call for further research. The importance of stability and loss functions in this emerging area emphasizes it's dynamic nature, which portends improved image captioning. In conclusion, the present proposal presents an overview on what the field is currently experiencing thus serving as a basis for more improvement and exploration in image captioning which is considered fascinating.

Keywords— Visual impairment, Long Short-Term Memory ,assistive technology, quality of life, blind individuals, machine learning, Recurrent Neural Network ,artificial intelligence, image captioning, neural networks, Convolutional Neural Network, real-time visual interpretation , Bilingual Evaluation Understudy.

1 Introduction

Captioning Image is a key process in the vast field of computer vision and natural language processing that allows visual content to be translated into meaningful texts. It's also a very complicated thing because it entails extracting relevant qualities from images, hence enabling machines to read between the lines of visual media [11, 35]. The bottom line is to capture the most significant attributes of an image, which can then be applied across multiple domains, like automatic news reporting, which is constantly changing [1]. At its core, image captioning unfolds through the lens of computer vision, where image encoding takes center stage. Additionally, this process goes into detail on objects in the image, their interrelationships, and larger scene information. Lastly, it provides a natural language caption that explains vital features in a picture, which opens up doors for novel applications in different domains [1]. Consequently, this complex procedure involves the detection of image features, the recognition of objects and their relationships, the understanding of scene information, and finally turning the visual content into meaningful sentences that are readable by humans. Color information as a primary data source has been integrated of late, marking the beginning of an era where convolutional neural networks (CNNs) and recurrent neural networks can work together for processing. Most researchers have instead been inclined towards employing neural network encoder-decoder architectures that are renowned for their impressive results in generating accurate and contextually rich descriptions (LIANG XU and WANSU LIM,2022).The LSTM networks also come into this story by providing coherence and linguistic support for the generation of such captions [15]. The panorama of image captioning methodologies unfolds with a myriad of techniques, including deep learning-based, retrieval-based and template-based approaches. The commonly embraced model involves an encoder-decoder sequence, cherished for its simplicity, flexibility in phrase structure, and adeptness in natural language annotation [24]. Variations in the form of two-stage and single-stage procedures have emerged, each addressing unique challenges within the image captioning framework [23]. An ingenious addition to this landscape is the Joint-Training Two-Stage (JTTS) technique, which harmonizes and collaboratively trains the tasks of the two phases, elevating the accuracy of the generated descriptions [22].

A sparse transformer-based approach for image captioning has emerged as a pressing need to ensure the efficiency of transforming pictures to text in a world that is overwhelmed with visual content. This can be achieved by reducing computation costs and increasing global context focus through attention mechanisms [28]. Nonetheless, it must also be noted that its efficacy depends on the specifics of datasets and particular task requirements, which bring their own challenges. In addition, the use of CNNs for this purpose is also presented in a variety of other research publications. Their model is well-constructed to include the extraction of features by CNN and the formation of linguistic descriptions made possible by LSTM (long short-term memory), which results in an impeccable mix between the two components [35]. Also, many published research papers support encoder-decoder architecture, where they utilize CNNs and RNNs to simplify image captioning while at the same time minimizing overall model size [36]. CNet-NIC has introduced a new framework that incorporates knowledge graphs to generate human-like and contextually rich captions. On the MS COCO dataset, experimental validation has been carried out that underscores how effective this innovative technique is and hence opens new ways of studying image captioning [37]. In addition, the I2CE metric takes advantage of word embeddings and auto-encoder principles to evaluate image captions, providing a much more nuanced perspective than traditional metrics [38]. The development of attention mechanisms plays a vital role in the progress of

image captioning models. The incorporation of decoders such as LSTMs or transformers with different types of encoders has made significant breakthroughs, especially in terms of efficient extraction of visual features and accurate phrase decoding [39]. Older models that did not include attention mechanisms struggled with arranging visual elements coherently and generating meaningful textual descriptions. Attention mechanisms have come up with remarkable improvements, although there still exist unresolved issues like capturing intra-alignments among items within an image [40]. Humans are endowed with this ability to decipher and describe visual situations in a snap. Humans instantly get information from visual stimuli without explicit explanations when reading news articles, browsing social media, or viewing ads. However, machines lack this intuition and often require direct instruction to decode visual data. Image captioning's main goal is to enable machines to generate authentic, linguistically coherent, and semantically meaningful descriptions of pictures that can help connecting the gap between visual comprehension and linguistic processing [41].

The primary goal of this literature review is not only to give an overview but also to narrate the story of the basic principles behind image captioning. It scrutinizes recent developments and illuminates the pros and cons of current models vis-à-vis extensive datasets [42]. Moreover, as we go deeper into this voyage, we grapple with unsettled issues in the field, which provide glimpses into ongoing difficulties and point out future research directions. The survey is not just a compilation of data but a reflection on the changing field of image captioning; hence, it is an invitation to both researchers and designers to join in the ongoing story of newness and perfection.

2 Methodology

Muhamad Zeeshan Khan et al.'s suggested work is divided into two sections: text encoding into semantic vectors and text decoding into natural pictures based on semantic properties. This design trains text and picture encoders at the same time using a fully trained generative adversarial network. Unlike previous approaches that rely on pre-trained text encoders, this method guarantees accurate image production through concurrent training. A convolutional neural network with three blocks is used for picture decoding, while a bidirectional long-short-term memory is used for text encoding. Two discriminators assess the characteristics and realism of the created pictures, and the generator contain an image decoder and a text encoder. For the assessment of face regions, the discriminator uses an attention mechanism. Both discriminator losses are included in the aggregate loss, guaranteeing adversarial training [1].

Extracting contextual information in image captioning involves capturing the relationship between visual features and the corresponding textual descriptions. Using attention mechanisms, which allow the model to concentrate on particular areas of the image when generating each word of the caption, is one method to accomplish this. This attention mechanism allows the model to dynamically adjust its focus based on the current word being generated, incorporating relevant visual context into the caption. Additionally, incorporating recurrent neural networks (RNNs) or transformer architectures enables the model to maintain contextual information across multiple words in the caption, ensuring coherence and relevance. By combining these techniques, the image captioning model can effectively extract and incorporate contextual information from the image into the generated captions.

"A novel framework for affective image captioning, influenced by models like M2, combines emotion attributes and cross-modal joint features. It employs affective tokens, encoder-decoder blocks, and gating/cross-attention mechanisms. The decoder utilizes masked self-attention and multi-level contributions for caption generation, along with emotion-based cross-attention mechanisms" (SHINTARO ISHIKAWA,2023). In another system, Hitesh Kandala et al. combine an LSTM-based

decoder for multilabel classification with a transformer-based encoder-decoder. Training optimizes the likelihood of generated captions given input images. The encoder employs a Transformer encoder for spatial information and CNN-based feature extraction using Inception v3. The multitask network utilizes label-smoothed cross-entropy loss for caption generation and binary cross-entropy loss for multilabel classification. Combining transformer and multilabel classification results in enhanced feature learning [3].

The encoder-decoder paradigm is widely used technique for image captioning, evidenced by the research conducted by CHUNLEI WU et al. The two different kinds of attention mechanisms are referred to as "hard attention" and "soft attention," respectively. “To enhance attention performance, network structures like CNN and LSTM are also extended. The neglect of low-level visual features aids in the comprehension of the images as well. Some suggest applying M-LSTM to interact with both textual and visual features to capture a high-level representation and using R-LSTM to identify which part of the captions are more important to the image. This makes it possible to include attention into a CNN that has an emotion polarity constraint. The LSTM language model serves as an agent and interacts with the word and visual contexts to train an example of this caption model. In addition, it is expected that the word "action" will appear next to the agent receiving the created phrase's score, or "reward," after the end-of-sequence (EOS) token has been established. As such, the agent has access to multi-grained incentives, or rewards, that rely on both sentence-level SN and REN. HAF and multicultural respect.” Consider implementing a fusion architecture of the attention model for captioning, which uses a multi-level feature map as input, as an alternative to employing a single image feature and concentrating on regions of the image with a single attention. Net Revaluation [4].

The decoder in the methods given by Huang Zhang et. al. [5]. creates words sequentially in a front-to-back sequence and is not capable of analyzing crucial contextual information. The Bi-LSTM (Bi-directional Long Short-Term Memory) structure used in this work gathers subsequent information in addition to prior information, allowing it to anticipate visual content based on context cues. As the hidden states are aligned and the semantic interaction is recovered based on similarity, the fused semantic information is the output. It uses a bi-LSTM-s model that can efficiently realize finer-grained picture captioning and extract contextual information.[5].

In the image captioning model by JU-WON BAE et al. [6], the architecture is divided into three domains: visual, decoder, and language. It employs an inject-based encoder-decoder architecture and utilizes a Geometry-Aware Self-Attention Network (GSA) to simulate geometric connections in pictures. Bi-LSTM and POS prediction are used to generate rich expressions, linking sentence information and pictures semantically. The decoder utilizes Bi-LSTM to create sentence information considering bidirectional context. The model embeds POS information and predicts POS using LSTM, enhancing caption creation [6]. The remarkable performance of deep learning models for real-time tasks like image classification, gesture recognition, video classification, natural language processing (NLP), instance segmentation, face recognition, and object identification has recently brought them greater attention. One of the most important jobs in NLP and computer vision (CV) is image captioning. This completes the image-to-text conversion process; more precisely, the model uses the input photos to automatically generate descriptive text. This paper builds a hybrid convolutional neural network image captioning system (LSAHCNN-ICS) for natural language processing (NLP) using the Lighting Search Algorithm (LSA). The LSAHCNN-ICS system, which was recently presented, creates an end-to-end model by using HCNN as the decoder and Shuffle Net, an encoder based on convolutional neural networks (CNNs). The Shuffle Net model extracts the image's feature descriptors during the encoding phase. Moreover, the hybrid convolutional neural network (HCNN) model that has been suggested can

be used to construct the text description during the decoding phase. The study's novelty, which aims to enhance captioning outcomes, involves applying the LSA as a hyperparameter tuning approach [7].

The WEITAO JIANG et al. system has “a self-attention mechanism equipped with a Multi-Gate Attention (MGA) block that extends the concept of conventional self-attention by adding an extra Attention Weight Gate (AWG) module and a Self-Gated (SG) module. The former limits the attention that should be given to the items that contribute the most. The latter is used to consider the distribution of intra-object attention and remove any unnecessary information from the object feature vector. Moreover, most existing picture captioning techniques directly enhance image attributes by using the original transformer created for natural language processing tasks. To streamline the transformer construction and increase its efficiency for enhancing picture features, they thus suggest a pre-layer norm transformer. They introduce a unique Multi-Gate Attention Network (MGAN) by integrating the AWG module into the language decoder and the MGA block with a pre-layer norm transformer architecture into the picture encoder.” (WEITAO JIANG).A pre-layer transformer is employed. Additionally, it illustrates the cutting edge of picture captioning. The SoftMax attention score is used by the attention weight gate module, which accepts queries, keys, and the sum of the attention weights as inputs. To remove unnecessary information from the object feature vector, it makes use of a proposed SG module that considers the intra-object attention distribution. Both the encoder and the decoder have MGAN [8].

In Tobias Hinz et. al.'s [9] proposed work, a generator, which creates new data points from randomly chosen inputs, and a discriminator, which attempts to discern between created and actual data samples, make up a typical generative adversarial network (GAN). Both the discriminator and the generator in conditional GANs are dependent on extra data. The AttnGAN is used as the foundational architecture. To enhance the quality of the produced pictures, AttnGAN, a conditional GAN for text-to-image synthesis, employs attention together with a unique extra loss. Three discriminators plus a generator make up this system. Image and caption similarity is calculated using the Deep Attentional Multimodal Similarity Model (DAMSM). During training, the generator receives extra, detailed data from this DAMSM on how well the created picture fits its caption.[9].

To enhance the performance of an image captioning network, several strategies can be employed. Firstly, leveraging pre-trained models like ResNet or VGG for image feature extraction provides a solid foundation. Fine-tuning these models on the specific dataset refines their understanding of image features. Incorporating attention mechanisms allows the network to focus on relevant parts of the image during caption generation, improving accuracy. Data augmentation techniques, such as rotation or cropping, expose the model to diverse perspectives, enhancing robustness. Lastly, continual evaluation, hyperparameter tuning, and incorporating human feedback ensure iterative refinement, ultimately leading to improved captioning performance

The image captioning model in the proposed work of DEEMA ABDAL HAFETH et al. uses an RNN decoder for language modeling and image caption construction, while a CNN encoder collects picture characteristics. using Long Short-Term Memory to decode visual information (LSTM). By using this strategy, the likelihood of the right description being given to the image is increased. One of the fundamental mechanisms for visual understanding is the object detection model. A quicker method for R-CNN object detection Sadly, there might be a lack of grounding in certain works when the object conceptions are unrelated to both the object locations and the description provided. These encourage the dissemination of semantics in visual attention throughout all suggested regions. They present a semantic-directed attention model that leverages concepts at the object instance level. The suggested work makes use of an image captioning knowledge base with common sense. To improve model

performance, this article focuses on combining common-sense knowledge bases with additional features and integrating them from outside sources. ConceptNet's extensive concept coverage and associated semantic embedding of helpful aspects make it a preferred external knowledge base when compared to a few other options. It makes use of multi-head attention and scaled dot product attention as an attention model.[10].

A bidirectional CARU (Bi-CARU) is presented in the work of Ka-Hou Chan et al. as a decoder for picture captioning jobs. This model also uses the attention mechanism to identify the characteristics for output encoding and the relationships between intriguing things in an image. With this design, context information can be efficiently extracted from decoding to produce prediction results that are more accurate. The section of speech that aids in aligning the hidden state extracted by the forward and backward CARU layers can be found via the context-adaptive [11].

LIANG XU et al.'s novel network model, *''bidirectional depth residuals gated recurrent unit network (BDR-GRU)"*, is designed and implemented to improve the effectiveness of image captioning in this paper. Initially, it makes use of the bidirectional network, which can be created using information from the past as well as the future. Second, a more intricate description is produced by the picture captioning through the usage of a deep structure. Lastly, the residual approach can generate a caption of higher quality while successfully preventing network degradation. [14]. The convolutional model used in this study by WANSU LIM et al. is constructed using an encoder-decoder architecture backed by a visual attention model. By using the pre-trained model from the PyTorch repository and the original convolutional architecture of Resnet, the encoder employs transfer learning. 1. With the use of a set of L-dimensional annotation/feature vectors, each of which represents a condensed representation of a different area of the original image, this procedure seeks to produce an encoded version of the input RGB image [15].

The suggested method, which investigates continuous outputs for language synthesis in the context of remote sensing picture captioning, is described by RITA RAMOS et al. Efficiency in the presence of enormous vocabularies was the primary driver behind the initial introduction of continuous outputs; nevertheless, we chose to concentrate on other possible benefits. Although continuous word representations, or embeddings, have been a common input for natural language processing (NLP) models, their application for language generating outputs has only been suggested in relation to machine translation recently. [20]

The early reliance on MS-COCO and other datasets created a solid foundation for training and evaluation in the field of picture captioning. The fundamental components of image captioning models are a decoder that generates coherent captions and an encoder that extracts visual data. Advancements in photo captioning models bring new methods. The integration of a visual scene graph offered by the Graph Attention Theory (GAT) model is a significant avenue. With the use of information from the scene graph, object detection is enhanced, resulting in more accurate and thorough descriptions of pictures.

1. Transformer-Based Methods:

Researchers increasingly leverage the Transformer framework to enhance caption generation, capitalizing on the superior performance of self-attention operations. "Through geometric self-attention, [29] developed an architecture that incorporates spatial relationships between things that are detected." (Herdade,2022). Within the Transformer paradigm, Li et al. [30] presented a novel attention mechanism that concurrently utilizes semantic and visual information.

2. LSTM Improved Methods:

Within the realm of sequence modeling, long-short-term memory (LSTM) plays a pivotal role in generating words based on input image features. Ke et al. [31] introduced a reflective position module and reflective attention module. While the latter decides how attention is distributed across input picture regions, the former pays attention just to hidden states.

3. Local Adaptive Threshold on Self-Attention:

Zhou Lei et al. [28] present a way of dealing with the efficiency of attention functions in the Transformer framework, more precisely for queries and several key-value pairs. Researchers have tried ways to build sparse attention matrices. Local adaptive thresholding on self-attention is introduced, which allows for more focused attention compared to the original Transformer framework.

The utilization of pre-trained vision-language models represents a promising step across various domains. Pre-trained algorithms, as demonstrated by the work of Itthisak Phueaksri et al. [21], showcase the potential of transfer learning in photo captioning. Emphasizing recognition of visual characteristics and meticulous pre-training on substantial corpora, including annotated object identification datasets, proves vital for enhancing model performance. Shifting focus to datasets, the MSCOCO Dataset emerges as a cornerstone for numerous computer vision tasks. With over 300,000 photos covering 80 object categories and five captions per image, it provides a rich resource. The typical train split comprises 82,785 photos, with 40,506 for validation and 40,773 for testing, following the 'Karpathy' data split. Additionally, the Flickr 30k dataset significantly contributes to dataset diversity, featuring 31,014 photos and 158,000 human-annotated captions. Researchers benefit from flexible training, validation, and test splits to tailor their choices according to specific research needs. Together, these datasets form a robust foundation for literature review and offer valuable insights into the evolving landscape of picture captioning.

Gaurav Joshi et al. suggested that the main part of their model is to provide a real-time description of an image. They utilized the Flickr 8K dataset to develop this project. In Flickr 8K, every image has five captions corresponding to it. The dataset provides 6000 images for training purposes, 1000 images for validation purposes, and the rest of the 1000 images for the 42 International Journal for Modern Trends in Science and Technology for testing purposes.[35]. To reduce the size of the reduced del in-image captioning architectures and increase performance, the goal is to choose the encoder model for image captioning. The faster R-CNN encoder model is used as an encoder, and without it, the modified version of Faster R-CNN MobileNetV3 is used for encoder compression to reduce the decoder model parameter without affecting the performance Pruning methods and quantization techniques have been used to reduce the size of the model without compromising its performance [36].

The study focuses on evaluating sentence similarity using vector semantics via Intrinsic Image Captioning Evaluation (I2CE). The approach is based on an auto-encoder framework that incorporates GRU units for phrase encoding and Bahdanau attention for decoding. Procedures for Semantic Representation: Uses pre-trained GloVe vectors to embed words and the NLTK library for stop-word removal to refine the semantic representation. Intrinsic vector extraction is accomplished using a GRU-based auto-encoder that fuses word vectors to produce a consolidated intrinsic vector expressing sentence meaning. The similarity measurement process compares the inherent vectors of candidate and reference captions, assessing their semantic closeness.[38]. The emphasis is on attention techniques within the framework of cutting-edge encoder-decoder systems for captioning images. For image caption generation, an attentive deep-learning model is used. Attentive deep learning combines computer vision, encoder-decoder architecture, and an attention method.[39]. In order to enhance the encoder-decoder framework's image captioning, the variational joint self-attention image caption model (VJSL) is employed. [40]

This technique first locates visually similar images with their captions in the training dataset, and the generated caption may be either an existing caption or a caption gathered from retrieved ones. Given a query image and descriptions retrieved from a set of existing descriptions or a set of predefined caption pools [41]. Introduce the attentive linear transformation (ALT) framework for automated picture caption creation. ALT Approach: Learn to pay attention to the heavy transformation function from visual feature space to context vector space. It can detect spatial focus, channel-wise recognition, and other important feature abstractions.[42]

Template-based methods used to extract important feature information from photos, such as objects, activities, sceneries, and sentences, using classifiers (such as SVM) apply predefined rules, n-gram dynamic fusion, lexical models, or templates to transform retrieved feature data into descriptive text. Visual Attention and Interactive Model Advancement makes use of mixed embeddings of main and secondary characteristics that are taken from picture salient areas, explaining how to use beam search and informative attributes to rerank caption possibilities as a way to enhance the model. adds more bottom-up elements to the Show, Attend, and Tell model's attention mechanism. attempts to surpass the most advanced picture captioning methods by cooperatively reordering beam search candidates.

Cross-Domain Research: Combines language processing with machine vision to improve picture captioning. focuses on improving model input for caption generation by utilizing textual attention in conjunction with visual attention methods. suggests utilizing textual attention in the Retrospect System for Image Captioning (RNIC) to enhance input and prediction processes. presents a textual attention technique to determine the words' contextual importance.[43]. The Third Study's Methodology Using an ensemble model, pictures are sent into an encoder to extract features, which are then fed into a decoder to produce captions depending on the information collected use state-of-the-art encoder-decoder architectures, such as sequential CNN-RNN, which are renowned for their capacity to generate outcomes in challenging tasks like voice synthesis. focuses on the difficulties associated with speech synthesis and feature extraction from various picture sequences [44].

3 Datasets & Components

In the study of Muhammad Zeeshan Khan et al. [1], numerous publicly accessible datasets, including face photos, are examined, such as “Celeb and LFW, which each have 11,000 images and provide information on gender, age range from young to old, hair and eye color, facial emotions, and ethnicity [1]. Artemis, which combines viewer emotions with visual art, is part of the SHINTARO ISHIKAWA et. al. dataset. Preprocessing involves using ResNet-32 and ImageNet classifiers to extract dominant emotion labels from images, which are then refined using ArtEmis. The ArtEmis dataset includes 455,683 emotional reactions and explanations for 80,031 WikiArt artworks. Its total vocabulary is 7,228,475 words, with an average sentence length of 14.9 words. Its vocabulary size is 37,250 words. The pieces span 45 genres (cityscape, landscape, portrait, still life, etc.) and 27 art styles (abstract, baroque, impressionism, etc.) from the 14th to the 19th century. 338,777, 19,931, and 39,850 samples make up each of the test sets [2], utilizing the experimental validation of the University of California (UC)-Merced captions dataset, which expands the widely-used UC-Merced dataset with class 21 and 100 images per class. large photos from the National Map Urban Area Imagery collection of the United States Geological Survey (USGS) for a variety of urban areas across the nation. This dataset has a pixel resolution of 0.3 m/pixel. Each image in the UC-Merced dataset has five reference sentences. The RSICD dataset does not apply to this system. It uses 80% of image captions as training data, 10% as validation data, and the remaining 10% as test data.” [3]

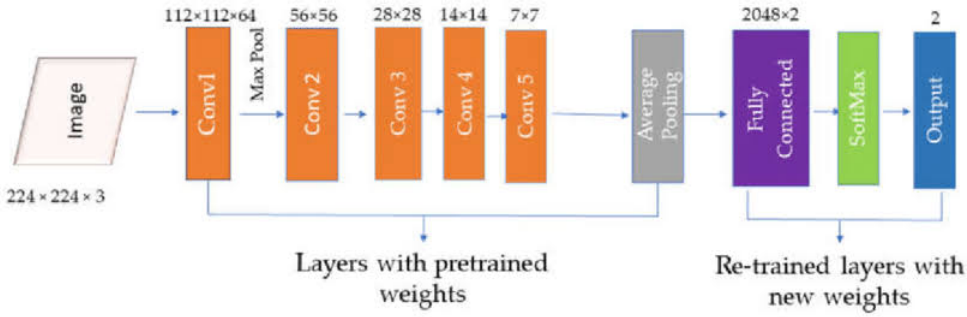

Figure 1: RESNET

MSCOCO 2014 datasets and an offline Karpathy split test have been employed. The split includes 5,000 photos for testing and validation and 113,286 training images with five captions each. The vocabulary is reduced to 9,680 words after filtering out any terms with fewer than five counts. the output of Resnet's convolutional layers conv4 and conv5, which map to a vector of dimension 1024 as the HAF's input. The LSTM hidden state, word embedding, and picture feature embedding dimensions for HAF are all set at 512. With an initial learning rate of 10−4, the baseline model is trained using the ADAM optimizer under the XE goal. Additionally, during every era, an offline Karpathy split test and MSCOCO 2014 datasets have been used. The split consists of 113,287 training photographs with five descriptions per image and 5,000 photos for testing and validation. After eliminating all words with a count of fewer than five, the vocabulary is reduced to 9,680 words. Convolutional layers conv4 and conv5, which map to a vector of dimension 1024 as the HAF's input, are produced by Resnet. For HAF, 512 is the value of the LSTM hidden state, word embedding, and image feature embedding dimensions. As usual, the baseline model is trained with an initial learning rate of 10−4 while utilizing the ADAM optimizer for the XE target. The reinforcement training starts at the 29th period in order to maximize the CIDEr metric with a learning rate of 10−5. The picture caption model is pre-trained with CIDEr rewards for 20 epochs during the word-level reward training phase. The hyperparameter margin α is set to 0.2 for the reward-level training, which lasts for 15 epochs.[4].

The main experimental dataset for image captioning is Zhang et al.'s images [5], which are gathered from everyday life. A multi-entity target with five manual labels is present in the single image to help with caption tagging. This dataset consists of 2.5 million labels, 328,000 pictures, and 91 targets. Eighty categories, more than 330,000 images—200,000 of which are annotated—and more than 1.5 million people are included in the biggest collection using semantic segmentation. 110,000 pictures are used for training, 5,000 for validation, and 4,000 for testing [5].

JU-WON BAE et al.'s model validation [6] involves training and testing on the Flicker 30K and MS COCO datasets. In Flicker 30K, 31,783 photos are split into 29,381 for training, 1,000 for validation, and 1,000 for testing, with 158,915 human-made description phrases. The MS COCO dataset follows Karpathy's split, with 123,287 photos used: 113,267 for training, 5,000 for validation, and 5,000 for testing, containing 616,435 sentences. The combined datasets feature 7,415 vocabulary entries, including special tokens. ResNetV2 serves as the intercepting model, using NLTK's POS tagger for POS data extraction [6]. For the LSAHCNN-ICS technique [7], a benchmark database is utilized for simulation analysis. It outperforms recent methods with maximum CIDEr scores of 43.61, 59.54, and 135.11 on Flickr8k, Flickr30k, and MS COCO datasets, respectively [7]. In addition, the MS COCO 2014 dataset is employed for training and evaluation, using the "Karpathy" split. It consists of 82,783

training images and 5,000 for testing and validation. Preprocessing involves lowercasing, tokenization, punctuation removal, and filtering terms appearing less than five times, resulting in 10,369 distinct terms [10].

In the study of Gaurav Joshi et al., numerous publicly accessible datasets are examined, such as Flickr 8K. In Flickr 8k, every image has five captions corresponding to it. The dataset provides 6000 images for training purposes, 1000 images for validation purposes, and the rest 1000 images for 42 International Journal for Modern Trends in Science and Technology testing purposes. Three main steps are involved: analyzing the data from the textual content; obtaining the distinctive vector from the picture; and deciphering the result by joining the many layers.[35]. We used the most popular picture caption benchmark dataset, MSCOCO, to examine the performance of several model-compression techniques. There are 40,502 validation photos and 82,885 training images in all. There are five different captions for each of the images. The framework that is used in this model is Python Torch, and the hardware used in the model is eight cores on the CPU and 32 GB of RAM, as well as one GPU accelerator for the Tesla V100. The models were trained for 30 epochs with the conventional method and then for an extra 10 epochs with the self-critical style.[36]

The benchmark data set for image captioning assessments that are most frequently used is the Microsoft COCO captioning data set (COCO). 40,504 photos are included in the validation set, and 82,763 images are used for training. Five or six descriptions or captions submitted by human annotators are included in the data set for every image. We employed 117,211 photos for training, 2,0 from the training and validation set that was supplied.[37]. The trials utilized the Microsoft Common Objects in Context (MSCOCO) dataset, with a focus on the MSCOCO-C5 branch. This branch includes five manual label phrases in each image to enable better generalization for assessment. The dataset also makes it easier to test the performance of hard matching and soft matching measures.[38]

We test the suggested approach using Flickr 30K and the MS-COCO dataset. The training and validation datasets make up the MS-COCO dataset. There are 123,487 photos in the training dataset, and each one has five label captions that have been manually annotated. There are 40,502 validation photos in the validation dataset. Every image in the MS-COCO collection has a caption consisting of five phrases. There are 40,502 validation images and 82,763 training images in the dataset, and each image has five label sentences. Thirty-seven hundred images and 158,925 sentences make up the Flickr30K dataset. Each image has five sentences connected with it. [41]. TensorFlow and Python (3.6 or above) are employed in the proposed product's backend development. Pandas will be used for data manipulation and cleaning, with Python loading and preprocessing the dataset. TensorFlow will build a deep-learning model trained on this data. This model is typically used on the front end of a web application to collect inputs and predict outcomes. The backend manages the seamless integration and processing of the dataset to ensure that it is compliant with the model. To meet TensorFlow's computing requirements, the hardware should ideally provide sufficient resources for model training and inference [44]. A sizable dataset called the Visual Genome is used to model how objects interact and relate to one another in an image. The 106K photos in the collection have detailed annotations of items, properties, and pairwise relationships. [45].

4 Evaluation Metrics

Muhammad Zeeshan Khan et al. [1] experimental research was conducted using a single Nvidia 1080Ti GPU with 11 GB of RAM and a GAN network. The model has an initial learning rate of 0.0001 and was trained for 500 epochs. The generator and both discriminators employ the Adam optimizer to maximize the weights. Since creating synthetic facial pictures that relate to real-world photographs is the aim of text-to-face synthesis, the distance between each feature in the two photos is calculated to

make the comparison. Face Semantic Gap (FSD) is the name given to this gap between facial characteristics. Additionally, they have contrasted the synthetic pictures' Frechet Inception Distance (FID) with the original images. First, 2048 inception characteristics from the pre-trained inception v3 model are computed to determine the FID. The system produces photos with a greater resolution, namely 256 × 256. the improvement in produced picture quality to the point that PSNR values, which range from 4.5 to 5, decrease as epochs grow.

Show Attend Tell (SAT) and M2, which have been successfully implemented for ArtEmis, are used by [2] as the baseline techniques. assessed using SPICE and CIDEr as the main metrics CIDEr received 15.4 points in the emotion-conditioned task, while the baseline approaches received 12.8 and 13.8 points. Furthermore, it scored 11.3 points in the grounded task, compared to 9.6 and 10.0 points for the baseline methods. For the two kinds of tasks, this method beat M2 on CIDEr by 1.6 and 1.3 points, respectively. Regarding SPICE, it received 8.3 points in the emotion-conditioned task, while the baseline approaches received 6.6 and 7.6 points. Furthermore, this approach scored 7.2 points in the grounded task, while the baseline methods scored 7.0 a points. Therefore, it outperformed M2 on SPICE by 0.8 and 0.6 points in the tasks.[2]

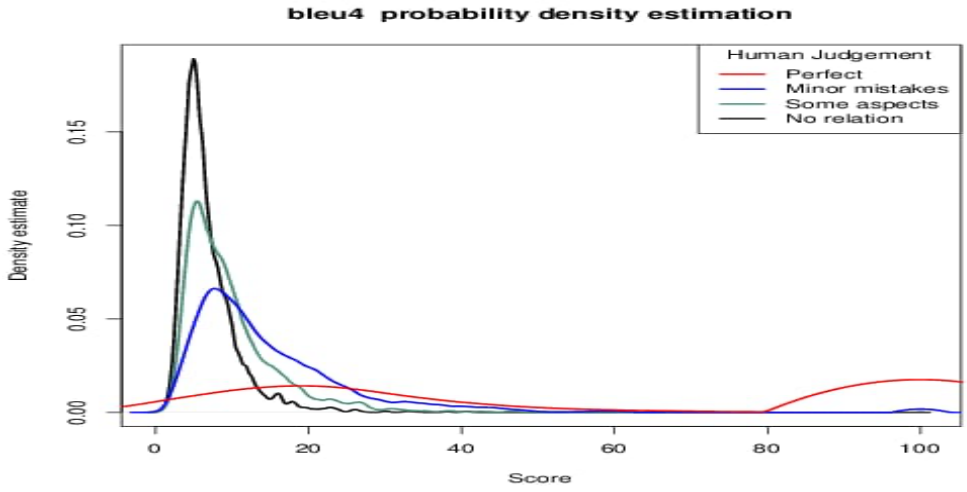

Figure 2 : Blue Probability Estimation

Results of reference [3] are shown using different popular indices: quality of the entire set of generated captions (BLEU) and its validity is compared with 4 checks; metric for evaluation of translation with explicit ordering (METEOR); "longest frequent subsequence understudy focused on recollection for gusting assessment (ROUGE-L); and consensus-based image description evaluation (CIDEr). LSTM (C + L) outperforms LSTM (C), showing that multilabel classification as an auxiliary task helps improve captioning, even for simpler LSTM-based architectures. Transformer (C) performs like LSTM (C + L). Efficiency decreases when simple auxiliary tasks are used. This shows that such auxiliary tasks are not consistent. When using multilabel classification as an auxiliary task, the proposed method performs significantly better than Transformer (C), LSTM (C), LSTM (C + L), and the scene attention-based technique. The LSTM in BLEU 1–4 is 0.766–0.547, METEOR is 0.388, ROUGE-L is 0.713, and CIDEr is 3.553. By contrast, the recommended set has values of 2.875 for CIDEr, 0.785 for ROUGE-L, 0.444 for METEOR, and 1 to 4 for BLEU [3]. The BLEU 1 for the reference [4] strong attention

approach is 78.1 and goes to 25.0 in BLEU 4, while the BLEU 1 for the soft attention method is 70.7 and goes to 23.9 in BLEU 4, with METEOR values of 23.0 and 23.9, respectively. The best performance is achieved when $\gamma = 15$ for the word-level reward via Ren HAF+REN, the scoring reward, and the CIDEr reward are balanced by varying the parameter β while assessing the effectiveness of sentence-level reward via SN. Their investigation reveals that $\beta = 0.3$ yields the greatest performance when considering {0.3, 0.5, 0.8} as the weight of the scoring reward. HAF produces superior descriptions than Topdown in several areas. This is the outcome of combining HAF (HAF+SCST) with multi-grain reward".([3] Sudipan Saha , Hitesh Kandala, Biplab Banerjee and Xiao Xiang Zhu,2022).

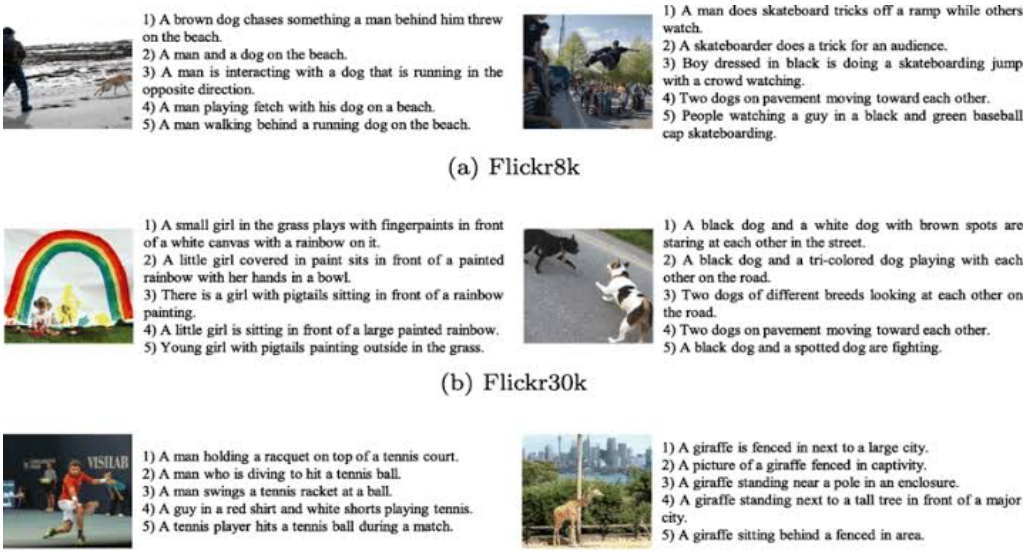

Figure 3: Flickr Typres

In image captioning, Long Short-Term Memory (LSTM) networks play a crucial role in generating descriptive captions for images. LSTMs process sequential information, making them suitable for understanding the context of images through their textual descriptions. They encode the visual features of an image and generate captions word by word, maintaining coherence and context throughout the caption generation process. LSTMs effectively handle the variability in sentence length and capture dependencies between words, enabling accurate and fluent caption generation for diverse images.

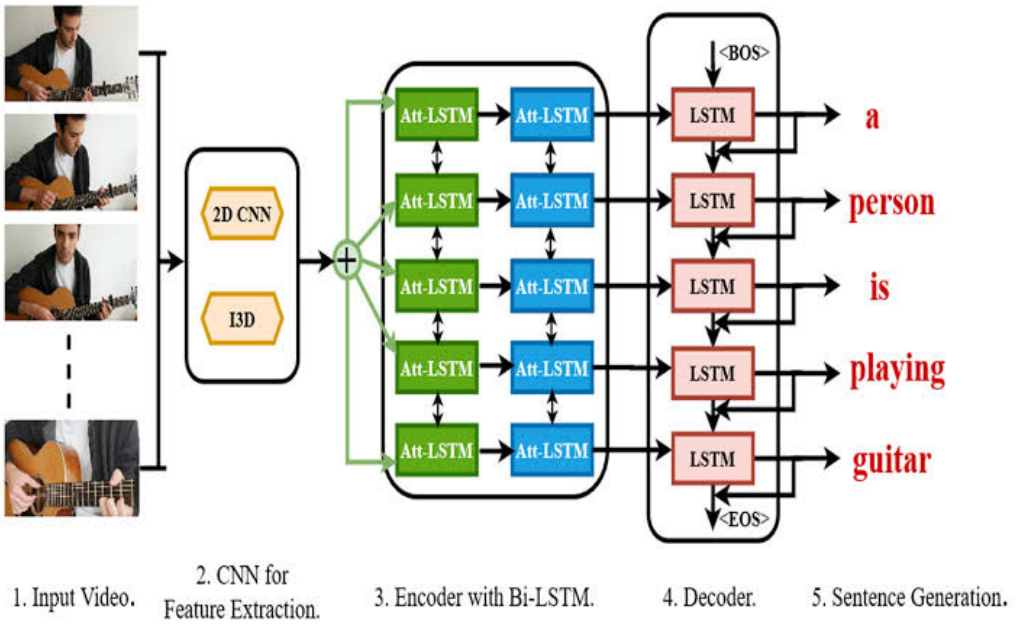

Figure 4: LSTM Network

Along with rouge-1 and spice, the likely utilized image captioning metrics are applied. "These were covered before. It is divided into two stages: the data preprocessing phase and the decoding phase. The two mechanisms are the parallel double-layer LSTM using MSCOCO and the auxiliary attention mechanism using MSCOCO. When the cross-entropy loss function is compared to the optimized CIDEr score (using the CIDEr score as an example), the leading methodologies currently in use significantly improve CIDEr based on cross-entropy error. Specifically, the CIDEr score of this Bi-LSTM model increased by 5.6 from 112.4 to 117.8, and the CIDEr score of their Bi-LSTM-S model increased by 2.7 from 118.6 to 121.3. Second, based on the aforementioned models, the Bi-LSTM-s improved from 112.4 to 128.6 by 6.1 in the cross-entropy loss function experiment and from 116.9 to 122.3 by 3.4 in the optimized CIDEr trial. pLSTM-A-2 concurrently encodes pictures using CNN and MIML, two different encoders." (HUAWEI ZHANG , CHENGBO MA , ZHANJUN JIANG, AND JING LIAN,2022).

The assessment measure for the JU-WON BAE et al. approach is TTR. "The total unique words divided by the total number of words in each phrase is known as the TTR. The suggested model's (LSTM) assessment criteria were raised from a minimum of 3.0 to a maximum of 11.1. The ROUGE-L score, for instance, increased by 11.1 points. The BLEU metrics, which compare linguistic manners simply, are particularly high. Over 80 in BLEU-1 and over 39 in BLEU-4 are evident. Furthermore, CIDEr uses TF-IDF in comparison to the n-gram captions. The highest CIDEr score is 131.2 for the M2 transformer. It categorized POS tags into nouns, verbs, ADJ, connective words, ADV, and DT; exclamations were included in the ETC category along with symbols and cardinal numerals." (JU-WON BAE,2021).

The Tobias Hinz et. al. [9] models perform better than the baseline AttnGAN in every metric. There is an improvement of 16–19% in the IS, 6-7% in the R-precision, 28–33% in the SOA-C, 22–25% in the SOA-I, 20–25% in the FID, and 15–18% in the CIDEr. computed each score using the COCO data set's

original pictures. They took 30,000 photos from the validation set, sampled them three times, and reduced them to 256 by 256 pixels for the IS. The CIDEr score was also computed using these pictures. They took three random samples of 30,000 photos from the training set and compared them to get the FID [9]. In the DEEMA ABDAL HAFETH et al. [10] study, BLEU@N, METEOR, ROUGE-L, and CIDEr-D metrics are used to assess an image captioning model. Semantic feature selection and ConceptNet embeddings are included in the proposed model, which outperforms state-of-the-art models in BLEU@1 (78.6%), ROUGE-l (57.7%), and CIDEr-D (120.98%). It does, however, trail somewhat in BLEU@4 and METEOR. The semantic attention of the model enhances relevance, appropriateness, and fluency. The model design performs optimally at eight heads, according to a quantitative assessment with different attention heads. The model's capacity to comprehend visual correlations is demonstrated through qualitative analysis, with semantic linkages improving picture descriptions. The suggested semantic-directed attention method's efficacy is validated through comparisons with a fundamental attention model.

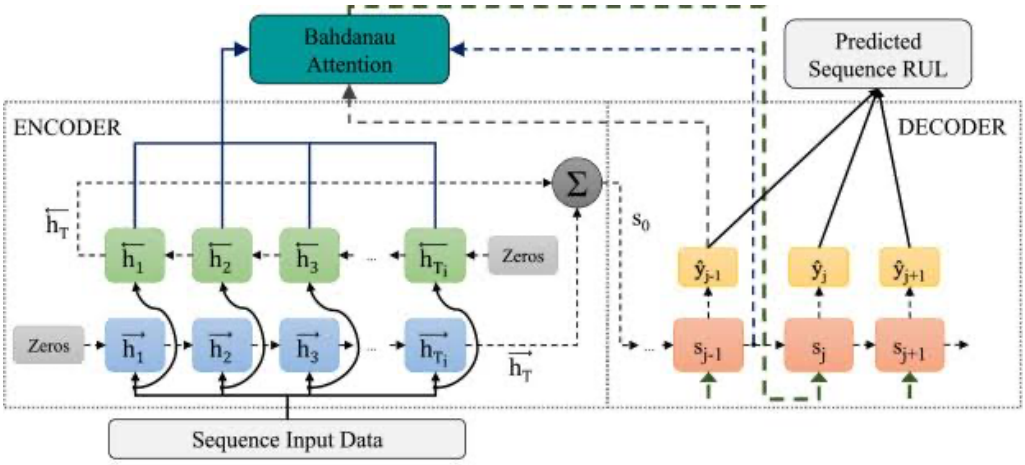

Figure 5: BAHD ANAU Attention Mechanism

Bahdanau attention improves translation accuracy by allowing models to focus on relevant parts of input text. It handles variable-length sequences effectively, enhancing performance in tasks like machine translation. The mechanism offers flexibility and interpretability, adjusting attention dynamically during decoding. It enables visualization of alignment between input and output sequences. Ultimately, Bahdanau attention boosts model performance in natural language processing tasks.

Ka-Hou Chang et al.'s validation of their work in the area of image captioning. We used the MSCOCO benchmark dataset in our CNN-to-Bi-CARU with other state-of-the-art works and two baseline methods. RDN, Show-Attend Tell, and four advanced methods (pLSTM, CNN BiLSTM-s, M3, and GRIT) are compared in our experiment. BLEU@5 and ROUGE-L were originally designed to evaluate machine translation, while CIDEr-D was specific. specifically designed to evaluate the accuracy of image descriptions against reference sentences [11]. The performance of VSAM is compared with that of five different models through two evaluation methods (detailed in Sec. IV.C.). The precision of VSAM is 0.112~0.126 better than the others when the evaluation method VKE I is used, while the recall of VSAM is comparable to the others. VSAM's precision, as measured by VKE II, is 91.7%, outperforming the other five models by 0.056 ~ 0.068. [12].

When we were analyzing the captions generated for sets of images, we used certain metrics that assessed both text-based and machine-learning components. For this purpose, we used BLEU [24], which is a measure of translation quality; ROUGE-L [25], which examines the quality of summary created; and CIDEr [26], which is an image captioning-specific metric. Additionally, we also employed some machine learning-based metrics that focus on semantic similarity between single words to turn particular words into more general concepts. According to findings made by Hyeryun Park et al. [27], among all metric scores, the mDiNAP-transformer-ewp model has been outperforming other models. The particularity of this model is its use of a transformer decoder and the inclusion of element-wise product feature difference vectors. This model differs from others in that it does not employ global average pooling. As the image captioning job gains popularity, there is a growing need for task-specific evaluation metrics that can be used to assess how well the created model is doing. Consequently, evaluation criteria such as CIDEr [33] and SPICE [34] developed. While models are capable of detecting many items, they may not always be able to effectively comprehend the relationships between them. For more precise and grammatically sound picture descriptions, using bigger datasets is recommended. Large-dataset training presents difficulties since it takes longer to train and test, which affects speed. Applications for those with visual impairments: By giving precise and audible descriptions of images, image captioning systems have the potential to greatly assist those with visual impairments in understanding their environment.[35]. It can be observed that a significantly reduced model size results in a comparably small change in performance metrics. [36]

By comparing n-grams—sequences of n words—in the candidate translation to those in the reference text, BLEU (Bilingual Evaluation Understudy) calculates how comparable the candidate translations and reference text are. It is taken from the assessment of machine translation. RECALL (Recall-Oriented Understudy for Analysis of Gisting): Originally intended for machine translation and summarization, it measures the overlap in n-grams between the candidate and reference texts to concentrate on recall. METEOR, or Metric for Evaluation of Translation with Explicit Ordering, is a metric that is computed using the harmonic mean of accuracy and recall for one gram. It prioritizes memory over precision by giving recall a larger weight than precision. A consensus-based image description evaluation tool called CIDEr. CIDEr was created especially for evaluating picture captions; it reflects human consensus better than other measures across sentences produced from different.[40]. To compare the actual and anticipated outputs, we employ the Bilingual Evaluation Understudy (BLEU) algorithm. This algorithm always produces an output that falls between 0 and 1. A number nearer 1 indicates that the texts are more comparable [42].

To evaluate the quality of image descriptions, the CIDEr (Consensus-based Image Description Evaluation) metric has gained a lot of popularity recently. It gauges how well the anticipated text matches the real label; a higher similarity suggests a more appropriate description and better prediction performance.[43]. The performance of several captioning models (such as RDN, Up-Down, and Att2in) on subsets of photos with variable average annotation lengths is compared in this statistic. Subsets with varying difficulties in scenarios This comparison evaluates the model's superiority, particularly in handling difficult circumstances, and validates the model's capacity to capture long-term relationships inside captions. When compared to other models that use typical LSTM or attention methods, this indicates the model's efficiency in handling complicated scenarios with lengthier annotations.[44].

5 Conclusion

This literature review has explored numerous approaches and difficulties related to computer vision in natural language processing within the broad field of image captioning research. It has become increasingly clear to us as we have explored the various facets of this multidisciplinary subject that to

create a compelling image caption, one must be aware of and in charge of every component of the image.

The exploration commenced with an overview of the CNN-to-Bi-CARU model. This model utilizes a bidirectional structure, integrating an attention mechanism, to enhance the extraction of contextual information from the context-adaptive features generated by the CARU's context-adaptive gate [11]. The objective is to facilitate a more comprehensive understanding and interpretation of image features, with potential applications in tasks such as image captioning. Image captioning is a multidisciplinary field bridging computer vision and natural language processing. Prior to translating visual content into understandable sentences, computers need to detect image features, identify objects, and determine relevant details. The integration of attention mechanisms is crucial for ensuring accurate alignment. However, traditional attention models are constructed within the decoder framework, leading to dynamic changes in focused content as words are generated. [12].

The analysis of inference time reveals that the tested images undergo efficient network calculations, demonstrating the model's suitability for real-time applications. Although the experiment's loss curve suggests some instability in the model, this article's scope does not delve into addressing this issue comprehensively. However, it is recommended for future research to explore and enhance the stability of the BDR-GRU model for further refinement [14]. During the initial experimental phase, it was observed that employing cross-entropy as the loss function yielded optimal results, achieving a Top-5 accuracy of 74.092 and a BLEU-4 metric of 0.201. Additionally, setting the Adam optimizer as an independent variable resulted in the best performance indicators, concluding the initial training phase with a loss value of 3.424, Top-5 accuracy of 74.092, and BLEU-4 score of 0.201 [15].

This study proposes an innovative framework for identifying significant regions within an image. The methodology involves leveraging image captioning to gather information from images and estimating crucial regions by associating them with subject and object words extracted from the generated captions. Through iterative training with a localizer, it was validated that the proposed approach can estimate significant regions with a sensitivity level closer to human perception [17]. Additionally, this paper presents a unique encoder-decoder model tailored for remote sensing image captioning. [20].

This survey has indeed uncovered various methodologies and insights that make image captioning a confluence of innovation and complexity. This field is ever-changing, from attention mechanisms to real-time efficiency matters to novel frameworks for region estimation, opening up avenues for improvement and exploration. The dynamic nature of this research domain is indicated by the challenges experienced in stability and loss functions.

To conclude, this survey provides an overview of today’s landscape while at the same time inviting researchers and practitioners to embark on a journey of exploration to address the identified challenges, refine methodologies, and usher in new waves of development in image captioning. There are possibilities for more refined, effective, and contextually rich image captioning models that blend the visual and linguistic worlds seamlessly.

Furthermore, enhancing the accuracy of image captioning models can be achieved through several strategies. Firstly, training on diverse datasets exposes models to a wider range of visual contexts, enabling them to generate more contextually relevant captions. Additionally, incorporating mechanisms such as the Bahdanau attention mechanism can further enhance model performance by allowing for more fine-grained alignment between image features and generated captions. By embracing these opportunities for refinement and innovation, there is potential to develop image captioning models that

are not only more accurate but also more effective and contextually rich, seamlessly bridging the visual and linguistic worlds.

References

[1] MUHAMMAD USMAN GHANI KHAN , MUHAMMAD ZEESHAN KHAN , SAIRA JABEEN , TANZILA SABA , ASIM REHMAT , "*A Realistic Image Generation of Face From Text Description Using the Fully Trained Generative Adversarial Networks"*,2020,IEEE Access.

[2] SHINTARO ISHIKAWA, *"Affective Image Captioning for Visual Artworks Using Emotion-Based Cross-Attention Mechanisms"*,2023,IEEE Access.

[3] Sudipan Saha , Xiao Xiang Zhu , Biplab Banerjee and Hitesh Kandala *"Exploring Transformer and Multilabel Classification for Remote Sensing Image Captioning",* 2022, IEEE.

[4] SHAOZU YUAN , YIWEI WEI ,CHUNLEI WU, AND LEIQUAN WANG "*Hierarchical Attention-Based Fusion for Image Caption With Multi-Grained Rewards*",2020,IEEE Access.

[5] HUAWEI ZHANG, ZHANJUN JIANG, CHENGBO MA AND JING LIAN, "*Image Caption Generation Using Contextual Information Fusion With Bi-LSTM-s*",2022,IEEE Access.

[6] JU-WON BAE , WON-YEOL KIM , JU-HYEON SEONG, SOO-HWAN LEE AND DONG-HOAN SEO "*Image Captioning Model Using Part-of-Speech Guidance Module for Description With Diverse Vocabulary*",2022,IEEE.

[7] SAMIA ALLAOUA CHELLOUG , RANA OTHMAN ALNASHWAN , NABIL SHARAF ALMALKI , IMENE ISSAOUI , AND AHMED SAYED " *Lighting Search Algorithm With Convolutional Neural Network-Based Image Captioning System for Natural Language Processing*" ,2023,IEEE.

[8] WEITAO JIANG , BOHONG, HAIFENG HU , QIANG LU, AND LIU XIYING LI "*Multi-Gate Attention Network for Image Captioning*" ,2021,IEEE Access.

[9] Stefan Heinrich, Tobias Hinz, and Stefan Wermter,"*Semantic Object Accuracy for Generative Text-to-Image Synthesis*",2022,IEEE.

[10] DEEMA ABDAL HAFETH , STEFANOS KOLLIAS, MUBEEN GHAFOOR "*Semantic Representations With Attention Networks for Boosting Image Captioning*",2023,IEEE Access.

[11] Ka-Hou Chan et al.'s " *Context-Adaptive-Based Image Captioning by Bi-CARU,2023, IEEE Access*.

[12] Yana Zhang et al.'s " *VSAM-Based Visual Keyword Generation for Image Caption,2021, IEEE Access*.

[13] YUEXIAN ZOU et al.'s *" Adaptive Curriculum Learning for Video Captioning,2022, IEEE Access*.

[14] LIANG XU et al.'s " *An Image Captioning Model Based on Bidirectional Depth Residuals and Its Application,2021, IEEE Access*.

[15] WANSU LIM et al.'s " *Deep Learning Approaches Based on Transformer Architectures for Image Captioning Tasks,2022, IEEE Access.*

[16] JIE FENG et al.'s " *Hybrid Attention Distribution and Factorized Embedding Matrix in Image Captioning,2020, IEEE Access.*

[17] SHINICHIRO OMACHI et al.'s " *Important Region Estimation Using Image Captioning,2022, IEEE Access.*

[18] WENYU CHANG et al.'s " *Skin Medical Image Captioning Using Multi-Label Classification and Siamese Network,2023, IEEE Access.*

[19] WEI YANG et al.'s " *Switching Text-Based Image Encoders for Captioning Images with Text,2023, IEEE Access*.

[20] BRUNO MARTINS et al.'s " *Using Neural Encoder-Decoder Models with Continuous Outputs for Remote Sensing Image Captioning,2022, IEEE Access.*

[21] Itthisak Phueaksri, Marc A. Kastner, Yasutomo Kawanishi, and Takahiro Komamizu, "*An Approach to Generate a Caption for an Image Collection Using Scene Graph Generation*," IEEE Access Digital Object Identifier 10.1109, Nov, 2023.

[22] Xiutiao Ye, Shuang Wang, Ruixuan Wang, and Licheng Jiao, "*A Joint-Training Two-Stage Method For Remote Sensing Image Captioning,*" Transactions on Geoscience and Remote Sensing, vol. 60, Nov. 2022.

[23] B. Wang, X. Zheng, B. Qu, and X. Lu, "*Retrieval topic recurrent memory network for remote sensing image captioning*," IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 13, pp. 256–270, 2020.

[24] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, ''*BLEU: A method for automatic evaluation of machine translation,*'' in Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL pp. 311–318,), 2001.

[25] L. Chin-Yew, ''*ROUGE: A package for automatic evaluation of sum- maries*,'' in Text Summarization Branches Out. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 74–81, Jul. 2004.

[26] R. Vedantam, C. L. Zitnick, and D. Parikh, ''*CIDEr: Consensus- based image description evaluation*,'' in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 4566–4575, Jun. 2015.

[27] Hyeryun Park, Kyungmo Kim, Seongkeun Park, and Jinwook Choi, "*Medical Image Captioning Model to Convey More Details: Methodological Comparison of Feature Difference Generation*," IEEE Access, vol. 9, pp. 1-1, 2021.

[28] Zhou Lei, Congcong Zhou, Shengbo Chen, Yiyong Huang, and Xianrui Liu, "*A Sparse Transformer-Based Approach for Image Captioning*," IEEE Access, vol. 8, pp. 190893-190904, 2020.

[29] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, ''*Image captioning: Transforming objects into words,*'' in Proc. Adv. Neural Inf. Process. Syst., pp. 11135–11145, 2019.

[30] G. Li, L. Zhu, P. Liu, and Y. Yang, ''*Entangled transformer for image cap- tioning,*'' in Proc. IEEE Int. Conf. Comput. Vis., pp. 8928–8937, Oct. 2019.

[31] L. Ke, W. Pei, R. Li, X. Shen, and Y.-W. Tai, ''*Reflective decoding network for image captioning*,'' in Proc. IEEE Int. Conf. Comput. Vis., pp. 8888–8897, Oct. 2019.

[32] A. Lavie and A. Agarwal, "*Meteor: An automatic metric for MT evaluation with high levels of correlation with human judgments*", Proc. 2nd Workshop Stat. Mach. Transl., pp. 228-231, 2007.

[33] V. Tech, C. L. Zitnick and D. Parikh, "*CIDEr: Consensus-based image description evaluation*", arXiv:1411.5726, 2014.

[34] P. Anderson, B. Fernando, M. Johnson and S. Gould, "*SPICE: Semantic propositional image caption evaluation*", Proc. 14th Eur. Conf., vol. 9909, pp. 382-398, Oct. 2016

[35] Gaurav Joshi1 Rd. Amita Goel2 | Vasudha Bahl3 Nidhi Sengar4 "*Image Captioning System*" ,2020, IJMTST

[36] Viktar Althia and Dmitrij Šešok "*Image-Captioning Model Compression*" 2022, MDPI

[37] Yimin Zhou, Yiwei Sun, Vasant Honavar " *Improving Image Captioning by Leveraging Knowledge Graphs*" 2019, arXiv:1901.08942v1

[38] Chao Zen*" Intrinsic Image Captioning Evaluation"* 2020 arXi:.2012.07333v1

[39] Zanyar Zohourianshahzadi* · Jugal K. Kalita," *Neural attention for image captioning review of outstanding methods*",2021arXiv:2111.15015v1

[40] Xiangjun Shao1 Zhenglong Xiang2,3 Yuanxiang Li1 Mingjie Zhang*," Variational join-self-attention for image captioning*",2022, IET

[41] Urja Bahety1, Surendra Gupta2" *Overview on Image Captioning Techniques*",2021, IJETER

[42] Vaquar Shaikh, Gaurav Verma, Soham Khade, R. A. Khan" *Image Captioning using Neural Networks Dhruvil Shah*",2022, IJARSCT

[43] Chaoyang Wang1, Ziwei Zhou1* and Liang Xu1" *An Integrative Review of Image Captioning Research"*, ISCME 2020

[44] Dhruvil Shah, Vaquar Shaikh, Gaurav Verma, Soham Khade, R. A. Khan "*Captioning Images using Machine Learning* ", 2022IJARSCT,

[45] Lei Ke1, Wenjie Pei2, Ruiyu Li2, Xiaoyong Shen2, Yu-Wing Tai2*," Reflective Decoding Network for Image Captioning",*