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

Generation of Image Captions using Deep Learning and Natural Language Processing: A Review

M Balakrishna Mallapu[1], Deepthi Godavarthi*[2]

Submitted: 28/01/2024 **Revised**: 06/03/2024 **Accepted**: 14/03/2024

Abstract: Deep Learning methodologies have significant possibilities for applications that endeavour to generate image captions or image descriptions automatically. Image captioning is among the most academically hard obstacles in image research. The caption of images is an extremely important study area that aims to automatically generate descriptive words based on an image's visual content. It's a multidisciplinary method that combines Artificial Intelligence (AI), Natural Language Processing (NLP), and Computer Vision (CV). Recognizing the Primary elements of the image, characteristics, and interactions is required for captioning. It should also generate sentences that are syntactically and semantically correct. Next, we evaluated the present literature discusses utilizing the language models to improve various applications, including image captioning, report creation, report categorization, extraction of findings, and visual query response and so on. In this article, we intend to present a comprehensive overview of available captioning of images using deep learning approaches. We also describe the datasets and assessment measures commonly utilized in deep learning for the automatic captioning of images.

Keywords: Deep learning, Natural Language Processing, Computer Vision.

1. Introduction

Image captioning is an intriguing exploration subject because of its numerous tasks, including assisting visually impaired persons, facilitating image categorization, as well as Natural Language Processing tasks [1]. Blind people rely heavily on image captions to access the internet daily. Simultaneously, recognizing images on the internet can be challenging for blind people [2]. The Internet is an invaluable resource for blind individuals, presenting them with the greatest degree of autonomy [3]. Blind individuals face challenges when it comes to accessing web data and carrying out routine activities like banking and grocery shopping etc [4].The captions of the images allow People with blindness to engage in social activities and get more information on the Internet, which aids in product purchases. Captions are generated automatically, allowing blind persons to learn more about the visuals. Captioning an image relates to the automated process of generating text that describes an image. In AI, creating descriptions for images receives growing attention and is becoming increasingly important [6].

Image captioning is a method that assists individuals in understanding media content and It emphasizes the most significant aspects that the recipient desires to convey in a visual representation [7].In general, image captioning tasks are classified into two categories: Natural Language Processing and Computer Vision [8]. In Computer Vision, image encoding is utilized to identify the objects within a frame and their interrelationships [9]. An NLP model receives the encoded feature and decodes it into a textual sentence [10]. The goal of image captioning is to provide natural language descriptions that accurately highlight the elements of incoming images.[11].The algorithm performs real-time analysis and deduction on the produced words and visual content while generating captions. [12]. An important challenge is managing the two distinct media formats during the encoding and decoding processes. The encoder-decoder strategy is used to address this difficulty, focusing on the entire surface of the image while creating captions [13]. In addition, the attention mechanism enhances the identification of noteworthy items by converting the sensory information becoming a set of attention weights or adjustable parameters, which are used for neurological training [14].

We commenced by elucidating the principles and historical background of language models, with a particular emphasis on expansive language models. Subsequently, we conducted a thorough examination of the existing literature about the utilization of language models in various applications including captioning of images, report preparation, Classifying reports, extracting findings, and responding to visual queries.

2. The Fundamentals of Language Models:

Models of languages are fascinating pieces of technology that mimic and understand human language. It analyzes

[1] *School of Computer Science and Engineering, VIT-AP University, Amaravati, Andhra Pradesh –522237, India*
ORCID ID :0009-0009-8526-0058
[2] *School of Computer Science and Engineering, VIT-AP University, Amaravati, Andhra Pradesh –522237, India*
ORCID ID :0000-0003-0712-6899
**Corresponding Author Email: deepthi.g@vitap.ac.in*

massive volumes of text data, identifying patterns and links between words, and then uses that information to anticipate the next word or series of words. Some of the applications include auto-completion, machine translation, chat bots, text summarization, content development, and speech recognition.

2.1 Recurrent Neural Networks (RNNs):

RNNs [16] (Rumelhart, Hinton, et al. 1985) are a kind of neural neural networks that are employed to examine successive inputs. RNNs, which are constructed from feed-forward networks, are distinct from conventional feed-forward neural networks in that they create directed cycles through their connections, enabling them to maintain a hidden state that stores details about previous sequence inputs. General designs for sequence learning problems include one-to-many, many-to-one, many-to-many (same), and many-to-many (different). This makes RNNs perfect for applications requiring sequential data, such as time series analysis, language modeling, machine translation, and speech recognition and captioning images. A significant challenge is the vanishing gradient problem, where gradients during training propagate backward in time and drop rapidly. As a result, RNNs have trouble determining long-term dependencies in sequences. Several sophisticated RNN variants, like GRU and LSTM networks, have been created. These architectures are better at handling long-term dependencies because they use gating systems, which permit the network to choose update, and Discard the information.

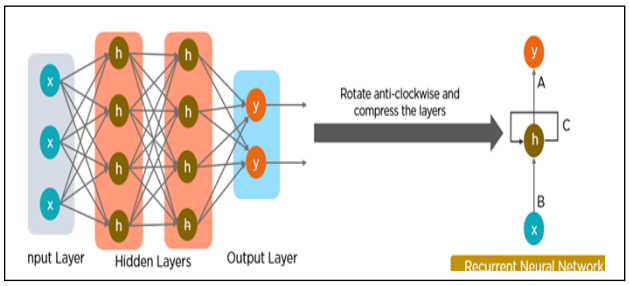

Fig. 1 The RNN's Architecture [16]

In the figure, x represents the input, the hidden state values represented by ‘h’, as well as the projected output as y. The network's parameters are A, B, and C. In RNN, the input at a given time stamp (t) consists of both the current input to time t and the output of the preceding of hidden state, h(t-1). Every timestamp results in an input loop.

2.2 Convolutional Neural Networks (CNNs):

CNNs[17] Utilizes deep learning techniques that accomplished substantial progress within the domains of image identification as well as classification. CNNs consist of multiple distinct fundamental layers, which are subsequently followed by activation layers. Convolutional neural networks consist of three layers, Pooling, fully connected, and convolution. The Convolution layer is a method in which a series of layers takes information from the input layer and extracts it at different levels. Concurrently, the fully linked layer creates and arranges the class label scores using the Softmax Classification algorithm. The pooling layer decreases the spatial dimensions among the complex features. There are actually two kinds of pooling, maximum pooling and average pooling. In maximize pooling, the maximum value is returned, while the former computes the average of all values from the picture within the kernel's boundaries. The Fully Connected (FC) layer provides classification according to the attributes acquired through their filters and the preceding layers. FC layers commonly classify inputs By utilizing a softmax activation function, which generates a probability ranging from 0 to 1. Several CNN architectures, including Google Net [23], Alex Net [19], Squeeze Net [21], ResNet [22], and VGG16 [20] have developed in recent years, with significant changes in hyper-parameters, layer types, and so on.

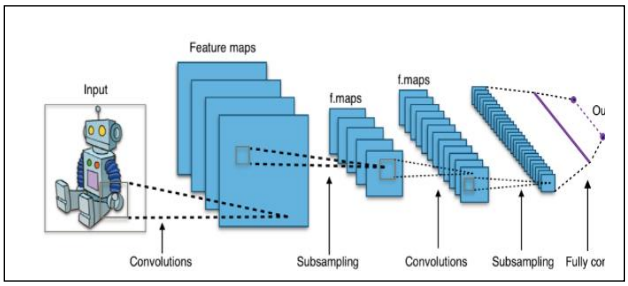

Fig. 2 Architecture of the CNN [17]

2.3 Long Short-Term Memory (LSTM)

Long Short-Term Memory networks [24] are a type of RNN that is specifically developed to handle a frequent RNN challenges the vanishing gradient problem. This difficulty happens when the network weights are modified using gradients during training become progressively less over time, making the network less able to understand long-term reliance in sequential input.

LSTMs address this problem by incorporating a unique cell structure featuring gates to regulate information flow, allowing them to selectively retain or forget data over lengthy periods. Cell State is at the center of the LSTM, transferring information across time steps. It functions as a conveyor belt, with gates controlling what goes in and out.

The input gate chooses the specific data to incorporate into the cell state from the available input at that moment. The Forget Gate determines which data from the prior cell state to discard, reducing unnecessary data

from cluttering memory. The output of the cell state is controlled by the output gate, which influences the network's predictions or actions. LSTMs are especially useful for sequence-based problems including speech recognition, Natural Language Processing, and time series prediction.

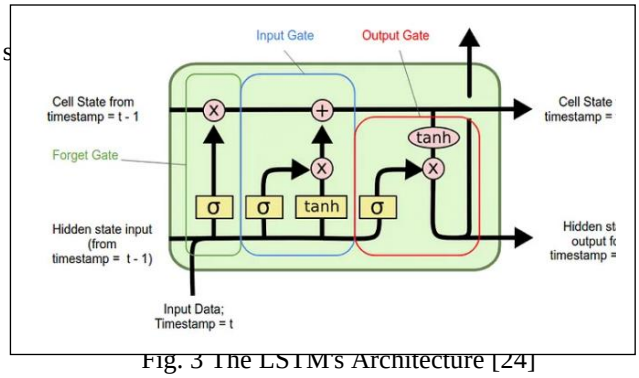

Fig. 3 The LSTM's Architecture [24]

2.4 Gated Recurrent Unit (GRU):

In deep learning, GRU [25], is a specific type of RNN architectural design. Kyunghyun Cho et al. introduced GRUs in 2014 as a simpler, more computationally efficient and alternative to LSTM networks. They perform well at tasks like sentiment analysis and machine translation when working with sequential data, dialogue systems, converting spoken language to text, and predicting future values based on previous data. GRU is intended to alleviate some of the shortcomings of classic RNNs, particularly in dealing with the vanishing gradient problem and handling long-range dependencies.

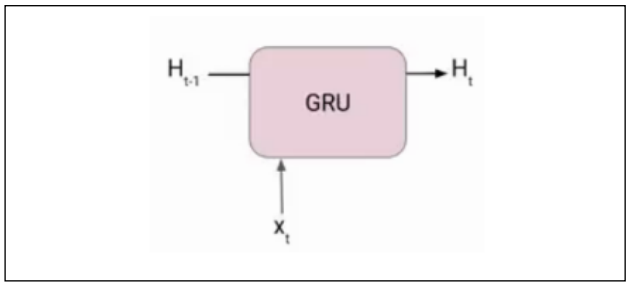

Fig. 4 The GRU's architectural design [25]

It takes as inputs Xt and the hidden state of the preceding timestamp, Ht-1. And then a novel hidden state of Ht is is brought back and delivered toward the subsequent timestamp.

Gates are used by the GRU to control the transmission of information throughout the network. GRUs have a hidden state that captures the network's memory during each time step. The current input and the previously concealed state are used to update this state. The Reset gate decides which data to delete from the prior concealed state, and the Update gate specifies the how much fresh data from the present input to incorporate into the updated hidden state.

2.5 TPGN:

"TPGN" stands for Tensor Product Generation Networks [26] are an interesting but less widely discussed approach to deep learning compared to LSTMs, GRUs, and Transformers. In 2017, TPGNs were originally presented as a novel architecture for applications related to Natural Language Processing (NLP), such as captioning images. They use tensor product representations (TPRs), a technique for distributed neural networks to process and encode symbol structures. TPRs encode relationships between elements in a sequence by multiplying vectors representing individual elements. This captures complex grammatical and semantic interactions within the sequence. The TPR-based approach allows them to explicitly encode and process grammatical structures, potentially leading to more grammatically correct outputs compared to other NLP models. The interpretability of TPGNs provides insights into their internal decision-making, valuable for research and debugging.

2.6 N-Gram:

According to Manning and Schutze (1999), an N-Gram [27] is based on literary order of n consecutive objects, which could be punctuation, words, numerals, or symbols. Many text analytics applications that use word sequences, such as sentiment analysis, text categorization, and text production, benefit from the use of n-gram models. N-grams are textual sequences made up of words, symbols, or tokens that are repeated. They can be described, technically speaking, as successive object sequences in a document. They are useful in NLP (Natural Language Processing) activities involving text data. They exhibit numerous uses, including language models, text mining, machine translation, semantic properties, and spelling correction. N-grams are classified into three types: unigrams, bigrams, and triangles. A bigram consists of two words, a trigram of three, and a unigram of one word.

2.7 Transformers:

Transformers are a neural network design that has become popular in the field of deep learning, particularly for applications related to Natural Language Processing. The design of Transformer had initially presented by Vaswani et al. in their 2017 publication "Attention is All You Need"[28].

Transformers, which resemble Recurrent Neural Networks or Long Short-Term Memory networks, use methods of self-attention rather than sequential processing to determine the value of unique words in a

series. Through this, they are better able to identify long-term dependence.

Transformers are usually employed in a sequence-to-sequence (seq2seq) configuration, where the input sequence is passed via an encoder, which then generates the output sequence using a decoder. The encoder and decoder layers of the Transformer are composed of feed-forward neural networks and self-attention mechanisms, respectively. The self-attention mechanism is usually combined with multipleattention heads to enhance the capability of the model to concentrate on distinct regions of the incoming sequence. Each head learns a unique linear projection from the input.

Natural Language Processing (NLP) has been dominated by transformers [28], with applications ranging from text-to-speech translation [31] to speech recognition [29], synthesis [30], and natural language production [32]. The transformer was the first deep learning architecture designed to address sequential inference challenges in Natural Language Processing. RNN [33] Utilize a series of inference methods, whereas transformers use layered self-attention mechanism is used to capture and retain long-term relationships in consecutive input. NLP applications often make use of transformer designs with large-scale topologies, such as BERT [35], GPT-3 [36], and the Transformer for Text-to-Text Transfer.

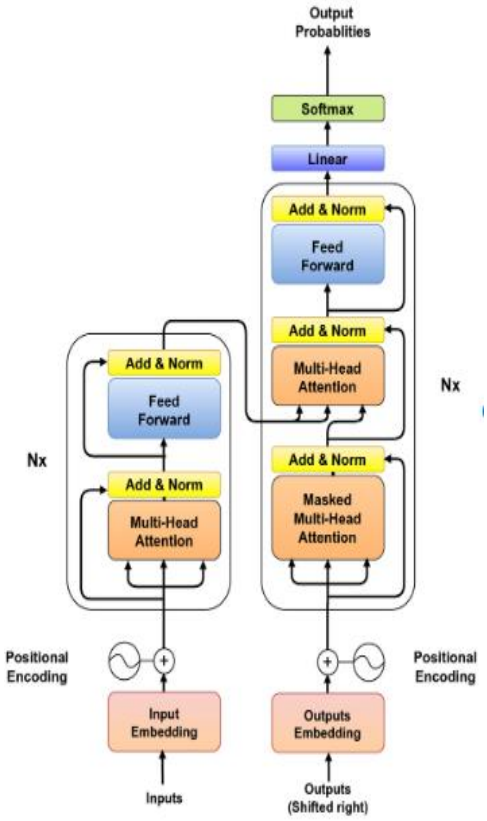

Fig. 5 The Transformer architecture [28]

3. The large-scale language models:

A massive language model can be described as an advanced form of linguistic systems that uses deep learning methods to train on enormous volumes of textual data. These models are capable of producing writing that resembles that of a human being and carrying out a variety of NLP functions.

Large Language Models are fundamental models for machine learning that analyze and comprehend natural language using deep learning methods. To find linguistic patterns and entity relationships, these models get trained using vast volumes of textual information. Many linguistic activities, including language translation, sentiment analysis and Chabot interactions, can be carried out by language learning machines. They can produce novel, coherent, and grammatically correct language, analyze complicated textual information, and identify objects and their relationships.

A deep learning method called a large language model is competent to manage a wide range of Natural Language Processing tasks. Large language systems are learned on vast datasets and utilize transformer models.

3.1 Bidirectional Encoder Representations from Transformers (BERT) :

BERT [38] (Lee and Toutanova 2018). A well-known large language model created by Google was trained using a substantial volume of textual input. It has the ability to comprehend a statement's context and respond coherently to inquiry. Transformer is a well-known attention model that is used for bidirectional language modeling training, which is the primary technological achievement of BERT. This is different from previous works that looked at a left-to-right text sequence or mixed training from left to right and from right to left. An attention mechanism called a Transformer is used by BERT to identify contextual connections within a text between words or sub words. Two independent systems comprise the transformer, a decoder that creates task predictions and an encoder that interprets the input text. It just needs the encoder mechanism because the objective of BERT is to build a linguistic model. The transformer translator reads each word in the string all at once. This makes it categorized as bidirectional. This feature allows the model to comprehend the meaning of a word by considering its complete surrounds, both to the right and left. In a variety of benchmarks, such as question answering, sentiment analysis, and natural language inference.

3.2. XLNet

This language learning module, developed by Google and Carnegie Mellon University, uses a novel method of

language modeling known as "permutation language modeling". It has achieved world-class performance in language tasks like sentence creation and question answering. By optimizing the expected probability across all possible permutations of the input text in a specific order of factorization. XLnet is an enhanced version of the Transformer-XL model which was pre-trained in bidirectional conditions using an autoregressive technique. It permutes the tokens in the phrase, allowing the model aims to forecast the next token (token n+1) based on the preceding n tokens. A permutation language model is called XLNet that differs from BERT in that its output predictions are generated in a random sequence. It looks at the encoded token pattern and predicts the tokens in a random sequence instead of a sequential one.

3.3. Large Language Model Meta AI (LLaMA) :

A high-level artificial intelligence (AI) system called LLaMA [39] is capable of understanding, creating, and analyzing human language. It is equipped with large-scale language models like GPT-3. LLaMA was provided in four different sizes (7B, 13B, 33B, and 65B parameters). Compared to earlier large language models, LLaMA creates text recursively by anticipating the word that comes after a string of words. OpenWebText, Common Crawl, and Wikipedia are just a few of the publicly accessible databases that provide the training data.

3.4 Parameterized Language Model (PaLM) :

Parameterized Language Model [40] Google's Pathways Language Model (Peng, Schwartz et al. 2019). PaLM was developed by OpenAI. The language model is autoregressive, that uses the context of previously created tokens to produce text by issuing tokens one by one similar to Generative Pre-trained Transformer models. Based on the preceding words' context, it forecaststhe probability distribution of the word that will come next in a sequence. Unlike traditional n-gram language models, PaLM can identify intricate patterns in data and manage long-term dependencies. The primary characteristic of PaLM is its capacity to manage words that are not included within the learning data, or out-of-vocabulary words (OOV). By replacing OOV words with contextually suitable ones, the model can effectively increase the overall efficiency of text creation.

3.5 Generative pre-trained transformers (GPT):

One of the most well-known large language models is the generative pre-trained transformer [41]. The popular fundamental model GPT was developed by OpenAI, and its iterations (GPT-3, GPT-4, etc.) have excelled their predecessors. In the future, it can be adjusted to carry out particular functions. Salesforce developed EinsteinGPT for CRM, whereas Bloomberg launched BloombergGPT for finance. It can perform a variety of jobs, such as text generation, interpretation, as well as summarization, and it has 175 billion parameters.. An artificial intelligence (AI) chatbot called ChatGPT mimics human-like conversational interaction by using natural language processing. In response to queries, the language model can produce a wide range of textual content, such as emails, articles, essays, social media postings, and code.

3.6. Text-to-Text Transfer Transformer (T5):

The Text-to-Text Transfer is an advanced architecture based on Transformers that can effectively handle various Natural Language Processing applications using an integrated text-to-text method. Instead of specializing in specific tasks like translation or question answering, it learns by translating all tasks into text formats.T5 relies on two main components, The Encoder Analyzes the input text and captures its meaning and the Decoder produces the desired result text according to the encoder's understanding. Both encoder and decoder utilize Transformer architecture, allowing for efficient parallel processing and capturing long-range dependencies in text.

3.7. Robustly Optimized BERT Pretraining Approach (RoBERTa)

RoBERTa was developed by Face book AI, an enhanced BERT variant that can handle a range of linguistic tasks. RoBERTa generally outperforms BERT on various NLP tasks, including General Language Understanding Benchmark (GLUE), SuperGLUE benchmark, Question Answering tasks

3.8. Bard:

Google AI developed a large language model (LLM), specifically the Gemini model. It is powered by LLMs' knowledge and capabilities but with a particular emphasis on conversational interactions. Google's Bard is an experimental conversational AI that runs on LaMDA (Language Model for Dialogue Applications), a Transformers-based conversational AI model that can be used to create dialogue-based applications. Bard creates fascinating discussions by utilizing massive datasets that comprise both textual and code data.

4. Literature Review

Image captioning can be categorized into three primary classifications: (1) Encoder-Decoder based image captioning, (2) Remote Sensing image captioning, and (3) Attention-based image captioning are reviewed in Table 4.1-4.3

Dai and colleagues (2017) explore the limits of current image captioning algorithms and present a novel strategy that employs Conditional Generative Adversarial

Networks (CGAN) to improve the generated caption's diversity and naturalness. The suggested strategy fared well in user studies and outperformed existing methods across a variety of tasks.

Gu and colleagues (2017) present a CNN model for language that is competitive in image captioning and well-suited for modeling statistical language applications. This model uses convolutional neural networks to detect long-range relationships in sequences, hence increasing the development of meaningful captions for images.

Wang and colleagues (2017) present a method for automatically producing descriptions of photos that are more close to how humans describe them. By dividing the description into a brief assertion and its characteristics, the approach can produce more accurate and innovative descriptions. The program also generates explanations of varying lengths to better suit user preferences.

Aneja et al. (2018) investigate the implementation of CNN for image indexing, tools for editing, and virtual assistants. The authors demonstrate that their convolutional image captioning technique is as effective as the classic LSTM approach but with a shorter training period.

Wang and Chan (2018) offer a new approach for automatically characterizing images that uses just convolutional neural networks, which is faster and produces better results than RNN or LSTM models. The trials reveal that the proposed CNN models dominate LSTM-based systems regarding performance in a range of metrics for evaluation.

Jie et al. (2021) address the issue of image captioning, where current approaches frequently produce generic and erroneous descriptions. The authors suggest a new method for producing more specific and detailed captions by employing a global-local discriminative aim. Their method surpasses existing approaches by producing captions that describe the images' visual details more accurately.

Zhou et al. (2021) provide a new network model dubbed BDR-GRU to improve the captioning of images. The design of the model is to operate on the mobile robot processor, it uses a BDR-GRU network to generate phrases and a convolutional neural network to encode images. Experimental results reveal that the BDR-GRU model beats other existing models on evaluation metrics.

Kastner et al. (2021) present a method for creating image captions that can be tailored to diverse purposes. The method allows you to regulate the level of length of the captions and their visual descriptiveness while keeping comparable captioning performance to other methods.

Tajrian et al. (2023) present a new approach known as the Vision Encoder-Decoder model, which comprises of interconnected models for encoding images and decoding text. They conducted research utilizing publicly accessible Bengali datasets and combined them to assess the effectiveness of their model, which generated more favorable outcomes than prior breakthroughs in Bengali image captioning.

Patil, Y., et al. (2023) present a model for producing visual descriptors using deep learning approaches. It uses a transformer architecture to create text sequences and a convolutional neural network for extracting visual characteristics. This model is trained using the Flickr8k dataset, which performs better than earlier methods.

Khustar Ansari. et al. (2024) introduce a deep learning model for automated optimization that generates captions for images.The system employs the encoder-decoder architecture, using the pre-trained ResNet 101 model to extract visual features and the SA-Bi-LSTM model for generating captions. A model of optimization known as the chip algorithm increases the performance of detection.

Akash Verma et al. (2024) introduces Encoder-decoder model using VGG16 Hybrid Places 1365 as encoder and LSTM as a decoder to generate accurate image captions achieving significant performance compared to existing approaches. The model is trained using the annotated Flickr8k and MS-COCO Captions datasets. The suggested method has demonstrated superior performance compared to existing state-of-the-art methodologies.

Hoxha et al. (2020) introduce a remote sensing image retrieval system that uses verbal descriptions to precisely define the connections between items and characteristics in images. The system encodes visual information, translates it into captions, turns the captions into meaningful feature vectors, and searches for comparable images based on vector similarity. Experimental findings reveal that this method produces accurate retrieval performance.

Huang et al. (2021) propose annotating remote-sensing photos by fusing multi-scale characteristics with a denoising approach. The suggested mechanism combines features of varying sizes and employs denoising during the visual feature extraction stage. This enables the encoder-decoder structure, widely used for image captioning, to represent denoised multi-scale features. In our studies, the method being suggested is to incorporate it inside the encoder-decoder framework and run comparative tests on UC Merced (UCM) and Sydney

captions are two freely accessible remote sensing image captioning datasets.

According to Li et al. (2021), In past RSIC frameworks, the model was trained using the cross-entropy function to anticipate the next word accurately. However, different sentences can be substituted with several synonyms. Consequently, over training occurs when the model has been learned, to anticipate a single word. In order to overcome this problem without over fitting, they propose a novel reduction cross-entropy objective function.

Ma et al. (2021) address the multi-scale problem by presenting two multiscale methods: multiscale attention and multifeat attention. These strategies seek to improve the representations Regarding the captioning problem in the remote sensing domain. The multi-head attention mechanism is employed with MSA technique to extract contextual information from data gathered from many layers. Similarly, The MFA method uses the target detection task to blend characteristics at the scene and target levels as an auxiliary work to improve the contextual feature.

Sumbul et al. (2021) introduce the novel summarization-driven RS image captioning technique. The proposed methodology consists of three fundamental components. The initial phase is to acquire standard picture captions using both CNN and LSTM. The subsequent stage is to combine each training image's genuine caption into a single caption by utilizing sequence-to-sequence neural networks, while also reducing any redundancy in the training set. Considering the semantic content of the image, the final phase automatically allocates adaptive weights to each remote-sensing image to integrate the conventional titles and the summary titles.

Wang et al. (2020) introduced are trieval RSIC method that splits the challenging RSIC task into two halves. The RS picture subjects are identified first. Secondly, a detailed depiction of the picture is constructed.

R. Ramos et al. (2022) investigate a novel technique for creating written explanations for aerial images by employing continuous outputs rather than discrete word tokens. The authors contend that this methodology can better capture the overall meaning of the captions, and experimental results suggest that it comprises outperforms the usual method on two datasets.

W. Nana et al. (2023) describe a pre-trained Bidirectional Encoder Representation of a Transformer that generates a contextually plentiful description embedding. The Transformer's Multi-Head Attention creates a strong association between the image and the contextually informed caption. We use the Dataset for Captioning Remote Sensing Images for this analysis.

Junsong Chen et al. (2024) introduce SMFE-Net, an end-to-end network using VGG-16 and LSTM for salient feature extraction and integration of high and low-level features. SMFE-Net utilizes an adaptive memory network to capture rich features, channel attention and spatial attention for adaptive feature extraction and achieves superior performance in VHR remote sensing image classification.

Wang et al. (2019) examine the ways that scene graphs, which serve as a representation of the semantic information included in images, can improve the captioning of images. The study concludes that a high-performing scene graph parser can greatly improve captioning accuracy, implying that the fundamental restriction is that instead of using the scene graph parser, use the captioning models.

Qin et al. (2019) introduce the techniques for enhancing picture captioning using Look Back and Predict Forward methods. LB makes use of historical visual data, whereas PF predicts the next two words at once. When combined, these strategies improve the functionality of various models for captioning images.

Wang et al. (2020) propose an innovative approach to accurately describing images. They use a recall mechanism to retrieve terms linked to the image and put them into the caption, resulting in better performance than other systems.

Cornia et al. (2020) provide M2, An innovative design that makes use of a model built on Transformers to improve image captioning. The M2 Transformer performs in the forefront of several tests and can characterize items that were not visible during training.

Zhang et al. (2021) introduced a Relationship-Sensitive model based on the Transformer for captioning images. It uses grid characteristics and relative geometry information to improve visual representations and a BERT-driven linguistic model with a module of adaptive attention to improve word prediction.

Luo et al. (2021) present a new network called the Dual-Level Collaborative Transformer which improves picture captioning by combining object recognition features with classic grid features. The DLCT network improves fine-grained details and contextual information by employing the cross-attention module and a distinct attention mechanism.

Fang et al. (2022) introduced a new image captioning model called ViTCAP that does not require a separate object detector. Instead, it predicts semantic ideas using grid representations and a Concept Token Network (CTN), resulting in a simpler architecture and

competitive performance in picture captioning challenges.

Wang et al. (2022) proposed an innovative method for annotating images which includes an architecture based on Transformers with a Swin Transformer backbone encoder. The model provides cutting-edge performance on picture captioning tasks without requiring pre-training on extra datasets.

Hafeth et al. (2023) offer a novel method called semantic attention network to improve image captioning. Combining general-purpose knowledge into the model, it improves the connections between items inside the picture and produces more specific and understandable captions. The Microsoft COCO dataset experiments show that our approach is contrasting with existing methods.

Himanshu et al. (2023) developed a novel algorithm called Local Relation Network (LRN) that can grasp an image's content and offer a natural language explanation. The method improves image representation and generates better captions by employing a multilayer attention technique and a form of Long-Short-Term Memory.

Bipul Hossen et al.(2024) proposed image caption generation using Guided Visual Attention (GVA) approach with two level attention mechanism for better captions. It utilizes LSTM for decoding and Faster R CNN for feature extraction and achieves significant improvements in caption quality on benchmark datasets.

Ravinder et al.(2024) proposed a Soft attention-based LSTM model that automates medical image captioning effectively by Using with YOLOv4 algorithm. It aims to address radiologists workload challenges and enhances description precision. The medical image content is automatically learned and described by the model by extracting information about objects and their special locations and generating decrypting sentences Using RNN and LSTM with attention mechanism.

4.1. A Table Featuring Image Captioning Techniques Based on Encoder-Decoder methods.

Ref	Year	Authors	Visual model	Language model	Optimizer
[56]	2017	Dai et al. (2017)	VGG Net	LSTM	-
[57]	2017	Gu et al. (2017)	VGG Net	Language CNN, LSTM	Adam
[58]	2017	Wang et al. (2017)	Res Net	LSTM	-
[59]	2018	Aneja et al. (2018)	VGG Net	Language CNN	RMSProp
[60]	2018	Wangand Chan (2018)	VGG Net	Language CNN	Adam
[61]	2020	Jie et al. (2021)	ResNet	LSTM	Adam
[62]	2021	Zhao et al. (2021)	VGG-16	Bi-GRU	Adam

[63]	2021	Kastner et al. (2021)	FasterR-CNN	BERT	-
[64]	2023	Tajrian et al.(2023)	ResNet	Transformer	Adam
[65]	2023	Patil,Y. et al.(2023)	EfficientNetV2B3, EfficientNetV2B0	Transformer	-
[66]	2024	Khustar Ansari et al. (2024)	ResNet 101	Bi-LSTM	Adam
[67]	2024	AkashVerma et al. (2024)	VGG 16	LSTM	Adam

4.2. A Table Including Techniques for Remote Sensing Image Captioning

Ref	Year	Authors	Visual model	Language model	Optimizer
[68]	2020	Wang et al. (2020)	VGG-16	LSTM	Adam
[69]	2020	Hoxha et al. (2020)	ResNet	LSTM	Adam
[70]	2021	Sumbul et al. (2021)	ResNet, DenseNet	LSTM	SGD
[71]	2021	Huang et al. (2021)	VGG-16, ResNet	LSTM	Adam
[72]	2021	Li et al. (2021)	VGG-16, AlexNet, GoogleNet	LSTM	Adam
[73]	2021	Ma et al. (2021)	ResNet	LSTM	Adam
[74]	2022	R.Ramoset et al. (2022)	EfficientNet	LSTM	-
[75]	2023	W.Nanalet et al. (2023)	VGG-16, VGG19	LSTM	Adam
[76]	2024	Junsong Chen et al. (2024)	VGG16	LSTM	Adam

4.3. A Table comparing different attention-based techniques for captioning images.

Ref	Year	Authors	Attention type	Image encoder	Image decoder	Optimization
[77]	2019	Wang et al. (2019)	Semantic	RCNN	LSTM	Adam
[78]	2019	Qin et al. (2019)	Hybrid	RCNN	LSTM	Adam
[79]	2020	Wang et al. (2020)	Hybrid	RCNN	Bi-LSTM	Adam
[80]	2020	Cornia et al. (2020)	Hybrid	RCNN, ResNet	Transformer	Adam
[81]	2021	Zhang et al. (2021)	Hybrid	ResNet	Transformer	Adam
[82]	2021	Luo et al. (2021)	Hybrid	RCNN	Transformer	Adam
[83]	2022	Wang et al. (2022)	Hybrid	Swin Transformer	Transformer	Adam
[84]	2022	Fang et al. (2022)	Hybrid	Vision Transformer	Transformer	Adam

[85]	2023	Hafethet al.(2023)	Semantic	CNN	LSTM	Adam
[86]	2023	Himanshu et al.(2023)	Multilevel attention	LRN	LSTM	Adam
[87]	2024	Bipul Hossen et al.(2024)	-	Faster RCNN	LSTM	Adam
[88]	2024	Ravinder et al.(2024)	Soft attention	RNN	LSTM	-

5.THE DATASETS AND EVALUATION METRICS

To train, test, and assess the image captioning systems, multiple datasets are used. There are significant differences in the datasets, include the quantity of images and the quantity of captions per image, their caption structure, and the proportions of the images. The three most widely used datasets are MS COCO Dataset [44], Flickr8k [42], and Flickr30k [43].

5.1 THE DATASETS

5.1.1. **Microsoft COCO Dataset.**

The Microsoft COCO dataset [44], A enormous collection used in many applications including identifying objects, segmentation of images, and captioning of images. It includes several characteristics, including picture segmentation, a large library of 328,000 photos, 91 different object classifications, and the unique feature of having five captions associated with each image.

5.1.2. Flickr 30K Dataset.

TheFlickr30K [43], This dataset allows for automatic visual description and contextual language interpretation.. There are 31000 Flickr images and 158k human-annotated captions included. For training, testing, or validation, it is unable to provide a stable image split. For training, testing, and validation, researchers can select the statistics they use. In addition, the collection has a preference for larger things, a classifier for colors, and monitors for common items.

5.1.3. Flickr8K Dataset.

Flickr8k [42] is a well-known collection that contais eight thousand pictures from Flickr. However, the test and development information each comprise 1,000 images, the training data contains 6000 images. Five reference texts provided by humans are included for each image in the collection.

5.1.4. Visual Genome Dataset.

As an additional option for image description, consider the Visual Genome dataset [45]. Reasoning about the relationships and features of the things in an image is just as important as just identifying them when captioning a picture. A section of an image is captioned by the Visual Genome dataset, in contrast to the other three groups. There are seven primary areas to the collection: pairs of questions and answers, descriptions of regions, objects, properties, connections, location graphs, and image graphs. Approximately 108,000 images make up the collection. A total of 35 objects, 26 attributes, and 21 pair wise interactions are present in each image on average.

5.1.5. Instagram dataset.

Tran et al. [46] and Park et al. [47] Utilizing visual representations from the social networking site Instagram, it produced two datasets. About 10,000 images make up Tran et al.'s collection, with celebrities making up the great majority of the images. On the other hand, Park and colleagues employed their dataset to predict hash tags and produce postings on social media platforms. This dataset includes 6.3 thousand individuals extensive hash tag list in addition to approximately 1.1 million posts with a wide range of themes.

5.1.6. IAPR TC-12 dataset.

The IAPR TC-12 dataset [48] contains 20,000 images. The pictures come from various sources, like pictures of sports, Humans as well as animals, and landscapes, among many more places all over the world. There are many language subtitles available for the photographs in this collection.

Table 4. The image captioning datasets

Ref	Year	Dataset Name	Images	Captions
[89]	2006	IAPR TC-12	20000	1-5
[90]	2013	Flickr 8K	8092	1-5
[91]	2014	Flickr 30K	31783	**5**
[92]	2014	MS-COCO	328000	**5**
[93]	2017	Visual Genome 1	108249	1-5
[94]	**2017**	Instagram dataset	10000	-

5.2 Evaluation Metrics

A number of metrics are employed to evaluate the effectiveness of systems for captioning images. These measures include comparing generated captions with reference captions using morphological similarity, contextual meaning, n-gram sequences, and other relevant characteristics. These measures are commonly used to assess image captioning.

- **BLEU(Bilingual Evaluation Understudy) :**

The BLEU [50], Papineni et al. (2002) measure is used to compare the similarity of generated and reference captions by calculating the precision of n-grams.. The degree to which the produced caption accurately matches the reference captions is measured.

- **METEOR(Metric for Evaluation of Translation with Explicit ORdering)**

The METEOR [51] , Banerjee, and Lavie (2005) propose an additional measure that uses n-gram precision but also takes into account recall and aligns words depending on their meanings. It considers synonyms and paraphrases, making it appropriate for evaluating various captions.

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

The ROUGE [52],].Lin (2004) proposes a metric that was originally designed to assess text summarizing activities but has now been adapted for image captioning evaluations. It uses n-gram analysis to assess the intersection between the produced and reference captions and the subsequence with the greatest number of similarities.

- **CIDEr (Consensus-based Image Description Evaluation):**

The CIDEr [53], Vedantam et al. (2015) determine the degree of agreement among the produced caption as well as the reference captions. It evaluates the degree to which the generated caption resembles each reference caption by considering the variety of human-provided titles.

- **SPICE: (Semantic Proportional Image Captioning Evaluation):**

The SPICE [54], developed by Anderson et al. (2016), is a measure utilized to assess the perfection of produced captions that determines how similar they produced and reference captions are based on semantic propositions.

6. Conclusion

Recent years have seen significant progress in image captioning. Accurate image captioning has significantly improved as a result of recent deep learning-based research. The efficacy of retrieval of images based on content can be enhanced by the textual description of the images, expanding the potential uses for visual comprehension in several industries, including the armed forces as well as medical applications etc. Simultaneously, the theoretical framework for image captioning and the research methods may result the advancements in picture annotation, cross-media retrieval, and visual question answering (VQA), with major academic and practical applications. This work will function as a catalyst for academics to create and innovate novel techniques to utilize linguistic prototypes to enhance captions for images.

References

[1] You Q, Jin H, Wang Z, Fang C, Luo J (2016) Image captioning with semantic attention. In Proceedings of the IEEE conference on computer vision and pattern recognition, 4651–4659

[2] Bigham JP, Lin I, Savage S (2017) The effects of not knowing what You Don't know on web accessibility for blind web users. In Proceedings of the 19th

International ACM SIGACCESS conference on computers and accessibility, 101-109

[3] Giraud S, Thérouanne P, Steiner DD (2018) Web accessibility: filtering redundant and irrelevant information improves website usability for blind users. International Journal of Human-Computer Studies 111:23–35

[4] Kuber R, Yu W, Strain P, Murphy E, McAllister G (2020) Assistive multimodal interfaces for improving web accessibility. UMBC Information Systems Department Collection

[5] MacLeod H, Bennett CL, Morris MR, Cutrell E (2017) Understanding blind people's experiences with computer-generated captions of social media images. In Proceedings of the 2017 CHI conference on human factors in computing systems, 5988-5999

[6] Bai S, An S (2018) A survey on automatic image caption generation. Neurocomputing 311:291–304

[7] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, and R. Cucchiara, ''From show to tell: A survey on deep learning-based image captioning,'' IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 1, pp. 539–559, Jan. 2023.

[8] T. Wolf et al., ''Transformers: State-of-the-art natural language processing,'' in Proc. Conf. Empirical Methods Natural Lang. Process., Syst. Demonstrations, 2020, pp. 38–45.

[9] X. Jiang, J. Ma, G. Xiao, Z. Shao, and X. Guo, ''A review of multimodal image matching: Methods and applications,'' Inf. Fusion, vol. 73, pp. 22–71, Sep. 2021.

[10] L. K. Allen, S. D. Creer, and M. C. Poulos, ''Natural language processing as a technique for conducting text-based research,'' Lang. Linguistics Compass, vol. 15, no. 7, Jul. 2021, Art. no. e12433.

[11] A. M. Rinaldi, C. Russo, and C. Tommasino, ''Automatic image captioning combining natural language processing and deep neural networks,'' Results Eng., vol. 18, Jun. 2023, Art. no. 101107.

[12] N. Xu, H. Zhang, A.-A. Liu, W. Nie, Y. Su, J. Nie, and Y. Zhang, ''Multilevel policy and reward-based deep reinforcement learning framework for image captioning,'' IEEE Trans. Multimedia, vol. 22, no. 5, pp. 1372–1383, May 2020.

[13] X. Xiao, L. Wang, K. Ding, S. Xiang, and C. Pan, ''Deep hierarchical encoder–decoder network for image captioning,'' IEEE Trans. Multimedia, vol. 21, no. 11, pp. 2942–2956, Nov. 2019.

[14] Z. Zohourianshahzadi and J. K. Kalita, ''Neural attention for image captioning: Review of outstanding methods,'' Artif. Intell. Rev., vol. 55, no. 5, pp. 3833–3862, Nov. 2021.

[15] S. Li, Z. Tao, K. Li, and Y. Fu, ''Visual to text: Survey of image and video captioning,'' IEEE Trans. Emerg. Topics Comput. Intell., vol. 3, no. 4, pp. 297–312, Aug. 2019.

[16] Rumelhart, D. E., G. E. Hinton and R. J. Williams (1985). Learning internal representations by error propagation, California Univ San Diego La Jolla Inst for Cognitive Science.

[17] LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. Proc. IEEE 1998, 86, 2278–2324. [CrossRef]

[18] Alom, M.Z.; Taha, T.M.; Yakopcic, C.; Westberg, S.; Sidike, P.; Nasrin, M.S.; Hasan, M.; Van Essen, B.C.; Awwal, A.A.; Asari, V.K. A state-of-the-art survey on deep learning theory and architectures. Electronics 2019, 8, 292. [CrossRef]

[19] Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. Commun. ACM 2017, 60, 84–90. [CrossRef]

[20] Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv 2014, arXiv:1409.1556

[21] Iandola, F.N.; Han, S.; Moskewicz, M.W.; Ashraf, K.; Dally, W.J.; Keutzer, K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. arXiv 2016, arXiv:1602.07360.

[22] He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778

[23] Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.

[24] H. Sharma and A. S. Jalal, ''Incorporating external knowledge for image captioning using CNN and LSTM,'' Mod. Phys. Lett. B, vol. 34, no. 28, Jul. 2020, Art. no. 2050315

[25] Cho, Kyunghyun, et al. "Learning phrase representations using RNN encoder-decoder for

statistical machine translation." arXiv preprint arXiv:1406.1078 (2014).

[26] Huang, Qiuyuan, et al. "Tensor product generation networks for deep NLP modeling." arXiv preprint arXiv:1709.09118 (2017).

[27] Manning, C. and H. Schutze (1999). Foundations of statistical natural language processing, MIT Press.

[28] Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need 2017. doi:10.48550/ARXIV.1706.03762.

[29] Dong L, Xu S, Xu B. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2018. p. 5884–8.

[30] Li N, Liu S, Liu Y, et al. Neural speech synthesis with transformer network. Proceedings of the AAAI Conference on Artificial Intelligence, vol 33; 2019. p. 6706–13.

[31] Vila LC, Escolano C, Fonollosa JA, et al. End-to-end speech translation with the transformer. In: IberSPEECH; 2018. p. 60–3.

[32] Topal MO, Bas A, van Heerden I. Exploring transformers in natural language generation: Gpt, bert, and xlnet. arXiv 2021:210208036.

[33] Graves A, Mohamed Ar, Hinton G. Speech recognition with deep recurrent neural networks. 2013 IEEE international conference on acoustics, speech and signal processing. IEEE; 2013. p. 6645–9.

[34] Sak H, Senior A, Beaufays F. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. arXiv 2014:14021128.

[35] Devlin J, Chang MW, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv 2018:181004805.

[36] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training; 2018. Available from https://s3-us-west-2. amazonaws.com/openai-assets/research-covers/language-unsupervised/language_ understanding_paper.pdf, 2018.

[37] Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. arXiv 2020:200514165.

[38] Lee, J. and K. Toutanova (2018). "Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805.

[39] Touvron, H., T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro and F. Azhar (2023). "Llama: Open and efficient foundation language models." arXiv preprint arXiv:2302.13971.

[40] Peng, H., R. Schwartz and N. A. Smith (2019). "PaLM: A hybrid parser and language model." arXiv preprint arXiv:1909.02134.

[41] Radford, A., K. Narasimhan, T. Salimans and I. Sutskever (2018). "Improving language understanding by generative pre-training."

[42] Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research 47 (2013), 853–899

[43] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In Proceedings of the IEEE international conference on computer vision. 2641–2649.

[44] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr DollÃąr, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In European conference on computer vision. Springer, 740–755.

[45] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision 123, 1 (2017), 32–73.

[46] Kenneth Tran, Xiaodong He, Lei Zhang, Jian Sun, Cornelia Carapcea, Chris Thrasher, Chris Buehler, and Chris Sienkiewicz. 2016. Rich image captioning in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 49–56.

[47] Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. 2017. Attend to You: Personalized Image Captioning with Context Sequence Memory Networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). 6432–6440.

[48] Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. 2006. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In International workshop ontoImage, Vol. 5. 10.

[49] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. 2011. Learning photographic global tonal adjustment with a database of input/output image pairs. In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 97–104.

[50] Papineni K, Roukos S, Ward T, Zhu W-Ji(2002) Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the association for computational linguistics, pp 311–318

[51] Banerjee S, Lavie A (2005) Meteor: an automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, pp 65–72

[52] Lin C-Y (2004) Rouge: a package for automatic evaluation of summaries. In: Text summarization branches out, pp 74–81

[53] Vedantam R, Zitnick CL, Parikh D (2015) Cider: consensus-based image description evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4566–4575, 2015

[54] Anderson P, Fernando B, Johnson M, Gould S (2016) Spice: Semantic propositional image caption evaluation. In: European conference on computer vision pp 382–398. Springer

[55] Kusner M, Sun Y, Kolkin N, Weinberger K (2015) From word embeddings to document distances. In: International conference on machine learning, pp 957–966. PMLR

[56] Dai B, Fidler S, Urtasun R, Lin D (2017) Towards diverse and natural image descriptions via a conditional GAN. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp 2989–2998, Venice. IEEE

[57] Gu J, Wang G, Cai J, Chen T (2017) An empirical study of language cnn for image captioning. In: 2017 IEEE International Conference on computer vision (ICCV), pp 1231–1240, Venice. IEEE

[58] Wang Y, Lin Z, Shen X, Cohen S, Cottrell GW (2017) Skeleton key: image captioning by skeleton attribute decomposition. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp 7378–7387, Honolulu. IEEE

[59] Aneja J, Deshpande A, Schwing AG (2018) Convolutional image captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5561–5570

[60] Wang Q, Chan AB (2018) CNN+CNN: convolutional decoders for image captioning. arXiv:1805.09019 [cs], May

[61] Jie W, Chen T, Hefeng W, Yang Z, Luo G, Lin Liang (2021) Fine-grained image captioning with globallocal discriminative objective. IEEE Trans Multimedia 23:2413–2427

[62] Zhao W, Xinxiao W, Luo J (2021) Cross-domain image captioning via cross-modal retrieval and model adaptation. IEEE Trans Image Process 30:1180–1192

[63] Kastner MA, Umemura K, Ide I, Kawanishi Y, Hirayama T, Doman Keisuke, Deguchi Daisuke, Murase Hiroshi, Satoh Shin'Ichi (2021) Imageability- and length-controllable image captioning. IEEE Access 9:162951–162961

[64] Ishan, Tajrian Islam, et al. "Bengali Image Captioning Using Vision Encoder-Decoder Model."

[65] Shetty, A., Kale, Y., Patil, Y. et al. Optimal transformers-based image captioning using beam search. Multimed Tools Appl (2023).

[66] Ansari, Khustar, and Priyanka Srivastava. "An efficient automated image caption generation by the encoder decoder model." *Multimedia Tools and Applications* (2024): 1-26.

[67] Verma, Akash, et al. "Automatic image caption generation using deep learning." *Multimedia Tools and Applications* 83.2 (2024): 5309-5325.

[68] Wang B, Zheng X, Bo Q, Xiaoqiang L (2020) Retrieval topic recurrent memory network for remote sensing image captioning. IEEE J Select Top Appl Earth Observ Remote Sens 13:256–270

[69] Hoxha G, Melgani F, Demir B (2020) Toward remote sensing image retrieval under a deep image captioning perspective. IEEE J Select Top Appl Earth Observ Remote Sens 13:4462–4475

[70] Sumbul G, Nayak S, Demir B (2021) SD-RSIC: summarization-driven deep remote sensing image captioning. IEEE Trans Geosci Remote Sens 59(8):6922–6934

[71] Huang Wei, Wang Qi, Li X (2021) Denoising-based multiscale feature fusion for remote sensing image captioning. IEEE Geosci Remote Sens Lett 18(3):436–440

[72] Li X, Zhang X, Huang W, Wang Q (2021) Truncation cross-entropy loss for remote sensing image captioning. IEEE Trans Geosci Remote Sens 59(6):5246–5257

[73] Ma X, Zhao R, Shi Z (2021) Multiscale methods for optical remote-sensing image captioning. IEEE Geosci Remote Sens Lett 18(11):2001–2005

[74] R. Ramos and B. Martins, "Using Neural Encoder-Decoder Models With Continuous Outputs for Remote Sensing Image Captioning," in *IEEE Access*, vol. 10, pp. 24852-24863, 2022, doi: 10.1109/ACCESS.2022.3151874.

[75] Nanal, Wrucha, and Mohammadreza Hajiarbabi. "Captioning Remote Sensing Images Using Transformer Architecture." 2023 International Conference on Artificial Intelligence in Information and Communication (ICAIIC). IEEE, 2023.

[76] Chen, Junsong, et al. "SMFE-Net: a saliency multi-feature extraction framework for VHR remote sensing image classification." *Multimedia Tools and Applications* 83.2 (2024): 3831-3854.

[77] Wang D, Beck D, Cohn T (2019) On the role of scene graphs in image captioning. In: Proceedings of the beyond vision and language: integrating real-world knowledge (LANTERN)

[78] Qin Y, Du J, Zhang Y, Lu H (2019) Look back and predict forward in image captioning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8367–8375

[79] Wang L, Bai Z, Zhang Y, Hongtao L (2020) Show, recall, and tell: image captioning with recall mechanism. Proc AAAI Conf ArtifIntell 34(07):12176–12183

[80] Cornia M, Stefanini M, Baraldi L, Cucchiara R (2020) Meshed-memory transformer for image captioning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10578–10587

[81] Zhang X, Sun X, Luo Y, Ji J, Zhou Y, Wu Y, Huang F, Ji R (2021) Rstnet: Captioning with adaptive attention on visual and non-visual words. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 15465–15474

[82] Luo Y, Ji J, Sun X, Cao L, Yongjian W, Huang F, Lin CW, Ji R (2021) Dual-level collaborative transformer for image captioning. Proc AAAI Conf ArtifIntell 35:2286–2293

[83] Wang Y, Jungang X, Sun Y (2022) End-to-end transformer-based model for image captioning. Proc AAAI Conf ArtifIntell 36:2585–2594

[84] Fang Z, Wang J, Hu X, Liang L, Gan Z, Wang L, Yang Y, Liu Z (2022) Injecting semantic concepts into end-to-end image captioning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 18009–18019

[85] D. A. Hafeth, S. Kollias and M. Ghafoor, "Semantic Representations With Attention Networks for Boosting Image Captioning," in *IEEE Access*, vol. 11, pp. 40230-40239, 2023, doi: 10.1109/ACCESS.2023.3268744.

[86] Sharma, Himanshu, and Swati Srivastava. "Multilevel attention and relation network-based image captioning model." Multimedia Tools and Applications 82.7 (2023): 10981-11003.

[87] Hossen, Md Bipul, et al. "GVA: guided visual attention approach for automatic image caption generation." *Multimedia Systems* 30.1 (2024): 50.

[88] Ravinder, Paspula, and Saravanan Srinivasan. "Automated Medical Image Captioning with Soft Attention-Based LSTM Model Utilizing YOLOv4 Algorithm." (2024).

[89] Grubinger M, Clough PM, Deselaers T (2006) The iapr tc-12 benchmark: a new evaluation resource for visual information systems. In: International workshop ontoImage, volume 2

[90] Hodosh M, Young P, Hockenmaier J (2013) Framing image description as a ranking task: data, models and evaluation metrics. J ArtifIntell Res 47:853–899

[91] Gong Y, Wang L, Hodosh M, Hockenmaier Julia, Lazebnik Svetlana (2014) Improving image-sentence embeddings using large weakly annotated photo collections. In: European conference on computer vision, pp 529–545. Springer

[92] Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: Common objects in context. In: European conference on computer vision, pp 740–755. Springer.

[93] Ranjay K, Yuke Z, Oliver G, Justin J, Kenji H, Joshua K, Stephanie C, Yannis K, Li-Jia L, Shamma DA et al (2017) Visual genome: connecting language and vision using crowdsourced dense image annotations. Int J Comput Vision 123(1):32–73

[94] Park CC, Kim B, Kim G (2017) Attend to you: personalized image captioning with context sequence memory networks. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp 6432–6440