IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Visual Analytics for Efficient Image-to-Text Prediction Based on Visually-Aware Context Learning

[1]NAGA VENKATA SUBRAMANYA NITHIN RANGA, [2]NANDIPATI VARSHITH NAGA SRI PAVAN,

[3]GUMMADI VARSHITHA,[4]S.REVATHY

[1,2,3]UG Scholar,[4] Assistant Professor Dept. of CSE, SRM Institute of Science and Technology, Ramapuram, Chennai

Abstract: *In the realm of image captioning, attention has predominantly been directed towards foreground objects, but a notable shift in focus has emerged, particularly evident in the context of geological images of rocks. Unlike traditional models that employ detection-based attention mechanisms, which often result in inaccurate captions by encompassing irrelevant backgrounds or overlapping regions, our approach seeks to address this challenge by refining attention to finer details. While convolutional neural networks (CNNs) have been a staple for both encoding and decoding in existing models, the crux of accurate image captioning lies in grasping the intricate semantic relationships between diverse objects within an image. Our methodology advances current practices by extracting feature vectors from meticulously segmented regions, enabling a more nuanced understanding of image components. Furthermore, we introduce a dual-attention module designed to independently process features from distinct classes, thereby enhancing the model's ability to discern complex scenes. Through rigorous experimentation, our model demonstrates proficiency in recognizing overlapping objects and comprehending scenes holistically, ultimately yielding competitive performance when benchmarked against state-of-the-art techniques.*

Keywords: Natural Language Processing, Ground Truth, Recurrent Neural Network, Bidirectional RNN, Region CNN.

I.Introduction

In our increasingly digital world, images have become ubiquitous across various technological devices, from smartphones to home security systems. With the surge in both real and artificially generated image data, the efficient storage, organization, and retrieval of images have become paramount. However, accurately retrieving images from a stored database poses significant challenges, as it requires systems to comprehend image details in a manner relevant to humans.

Image captioning, a focal point of computer vision technology, plays a pivotal role in scene understanding by detecting objects, discerning their relationships, and describing semantic content using natural language. Traditionally, image-captioning methods have translated extracted image features into descriptive text, offering simplistic scene descriptions. However, recent advancements in computer vision and deep learning have shifted the focus toward achieving precise and comprehensive image captioning.

Most people acknowledge vision as the primary sense for perception, with the human brain being adept at processing visual information. Indeed, visual data are processed more efficiently than any other form of information, with the brain rapidly analyzing and contextualizing images upon perception.

Image captioning tasks involve generating textual descriptions of image semantics, with a key challenge

being the production of distinctive captions that uniquely identify images. Unlike generic captions, distinctive ones offer more informative and descriptive insights, making them valuable for retrieval applications and aiding individuals with visual impairments.

Standard image captioning datasets typically describe salient objects in images, often resulting in generic captions shared across similar images. To address this issue, modern deep learning-based methods leverage encoder-decoder frameworks, comprising a Convolutional Neural Network (CNN) for image feature extraction and a Long Short-Term Memory (LSTM) model for caption generation.

Image captioning represents a complex multimodal task bridging computer vision and natural language processing. This involves comprehending visual content, identifying salient elements, and accurately describing them using natural language. While traditional models primarily focus on global image regions, recent advancements in attention mechanisms, such as visual and semantic attention, have significantly enhanced interpretability and performance in image captioning.

II. LITERATURE REVIEW

Recent advancements in remote sensing image captioning often overlook the disparities between remote sensing and natural images. To address this, we propose a multiscale multi interaction model that adapts to remote sensing image characteristics. Our model incorporates a two-stage multiscale structure for feature representation and a multi interaction module to enhance feature distinguishability. Experiments on various datasets demonstrate significant improvements across multiple evaluation metrics, showcasing the efficacy of our approach. [1].

The study introduces a novel framework for identifying significant regions within images by leveraging image-captioning techniques to interpret contextual information. This method aims to identify important regions more accurately than conventional saliency-based approaches. A dataset was created to define these regions based on subjective evaluations. The proposed approach outperformed traditional methods in terms of accuracy. By utilizing semantic information from image captions, the method identifies regions corresponding to subject and object words, achieving results closer to human perception. Future work includes addressing captioning failures and improving attention accuracy. [2].

The study introduces a novel framework for identifying significant regions within images by leveraging image-captioning techniques to interpret contextual information. This method aims to identify important regions more accurately than conventional saliency-based approaches. A dataset was created to define these regions based on subjective evaluations. The proposed approach outperformed traditional methods in terms of accuracy. By utilizing semantic information from image captions, the method identifies regions corresponding to subject and object words, achieving results closer to human perception. Future work includes addressing captioning failures and improving attention accuracy.[3].

The Visual Semantic Attention Model (VSAM) seems to be a key component of your approach, achieving an impressive precision of 91.7% in visual keyword generation. This high precision suggests that VSAM is effective in identifying and extracting relevant visual cues from images. By improving the accuracy of image captioning through better alignment of visual and semantic information, your work could have significant implications for various applications, including image search, assistive technologies, and content creation. It'll be interesting to see how this concept of visual keywords evolves and contributes to advancing the field further![4].

A novel chest x-ray image captioning model designed to automatically generate draft reports, easing the burden on doctors. By accounting for differences between patient and normal images and exploring various feature representation methods, including hierarchical LSTM and transformer decoders, your approach shows promise in accurately interpreting medical images. Comparative analysis against recent captioning approaches, using metrics like BLEU, METEOR, ROUGE-L, and CIDEr, identifies the multi-difference non-average-pooling transformer model as the top performer, affirming its effectiveness in generating draft reports. This model not only offers practical utility for medical professionals by saving time and reducing expenses but also hints at broader applicability to other medical image types, suggesting avenues for further research and development in medical image captioning. [5].

The current issues in picture captioning are related to generating the captions which have low semantic dissimilarity against the information transmitted by the image and have high syntactic readability. Thus, in the attempt to mitigate these challenges, a novel method of picture captioning, namely ATT-BM-SOM, is proposed. The new approach to image captioning consists of attention checking mechanism and syntax optimisation module. It successfully merges visual data and generates high-quality captions by selecting the appropriate image

features and optimising the syntactic structure of the captions. The model's outstanding performance is shown by the experimental results on the MS COCO dataset, which indicate its high scores on BLUE-1, ROUGE-L, CIDER, and SPICE, respectively. The model's ability to generate captions that are easier to read and provide detailed explanations is supported by both quantitative and qualitative data, setting it apart from existing baseline methods.[6]

We propose a novel deep encoder-decoder model for image captioning based on the sparse Transformer framework. Our model effectively captures correlations between image regions and words using self-attention mechanisms. Additionally, we introduce a Local Adaptive Threshold mechanism to enhance attention concentration, improving word generation accuracy. Experimental results on MSCOCO and Flickr30k datasets demonstrate superior performance compared to previous methods[7].

The proposed procedure involves several key steps to improve image semantic segmentation, particularly in remote sensing applications. These steps include dividing images into multiscale features, restructuring the deep learning network model, jointly predicting across multiple scales, and optimizing postprocessing using a fully connected conditional random field (CRF). Inspired by scale-space theory, hierarchical multiscale division processing is implemented on images to capture detailed information at different resolutions. Furthermore, the architecture of the DeepLabV3+ model is enhanced, and the feature output layer is modified to incorporate multiscale features through weighted fusion, aiming to enhance segmentation accuracy and robustness. [8].

This research presents a new cascade semantic fusion architecture (CSF) that utilizes various forms of semantic information extracted from pictures to improve the process of generating image captions. The CSF is specifically developed to gather comprehensive object information and contextual information around objects using a three-stage cascade procedure. During the first phase, object-level attention characteristics are obtained by using a pre-trained detector. In the second step, the object-level characteristics are combined with spatial information to create image-level attention features, which enhance the surrounding context of the objects. During the last phase, spatial attention characteristics are acquired to enhance the attention features that were previously taught. The CSF effectively organizes contextual information about pictures from various viewpoints by incorporating attention processes with these three sorts of properties. The CSF has been shown to be successful in picking object areas of interest and producing more accurate picture captions using the MSCOCO dataset. It outperforms numerous current image captioning systems in terms of performance. [9].

The advent of image captioning technology has transformed our capacity to comprehend and articulate the content of images via the use of machine intelligence. The research focus has shifted towards using deep learning to understand visual information and generate descriptive text. This study presents a new method called the multilayer dense attention model for picture captioning. The approach we use involves the use of a Faster R-CNN to efficiently extract picture characteristics, which acts as the coding layer. The decoding technique utilizes LSTM-Attend to unravel the intricate multilayer dense attention model and provide descriptive text. Parameter optimization in reinforcement learning is accomplished via strategy gradient optimization. Our approach efficiently filters out irrelevant information by integrating dense attention processes into the coding layer. This allows for the selective generation of relevant descriptive text during decoding. The model's ability to understand images and generate text has been confirmed by experiments done on several picture datasets. We propose our results as a substantial addition to the area of picture captioning. We would like to express our gratitude to the reviewers for their invaluable comments. [10].

III.METHODOLOGY

The methodology employed in this study adopts a multifaceted approach to enhance the performance of image captioning. It begins by extracting region-level features, where target-detection techniques are employed to detect object relationships and construct a graph structure. Utilizing Graph Convolutional Neural Networks (GCN), region-level features are extracted to guide the LSTM in generating captions, ensuring a comprehensive understanding of the image content. Following this, the Domain Object Pre-Filtering process adjusts the order of image objects and region information input to the captioning model based on a domain object dictionary. This step ensures the prioritization of specific objects, enhancing the relevance and accuracy of the generated captions. Moving to Text Generation, the study addresses challenges in natural language processing by employing techniques such as top-k sampling to enhance the diversity and creativity in text generation. The Visually-Aware Context Network, a critical component of the methodology, focuses on enhancing semantic representations of text using various techniques, including residual networks and atrous convolution. These methods enrich the semantic content of captions, resulting in more informative and contextually relevant descriptions of the image content.

Finally, the methodology includes the Evaluation Metrics module, which assesses captioning performance using standard metrics such as BLEU, METEOR, ROUGE, CIDEr, and SPICE. This comprehensive evaluation provides insights into the effectiveness of the proposed approach in improving the accuracy, diversity, and semantic richness of image captions, thereby advancing the state-of-the-art in image captioning technology.

IV. DISCUSSIONS

A.Importance of Image Captioning

The ubiquity of digital images in contemporary technology, spanning smartphones, computers, and various smart devices equipped with cameras, underscores the critical role of image captioning. This technology has become increasingly indispensable with the emergence of Vision-Language Models, which possess the remarkable ability to generate high-fidelity images from textual descriptions. As a result, the need to efficiently organize and retrieve image data has become more pronounced. Image captioning serves as a pivotal link between the realms of computer vision and natural language processing, empowering machines to comprehend and articulate visual content in a manner akin to human perception. Its significance extends across diverse domains, including aiding visually impaired individuals in interpreting visual information, enhancing the search capabilities of engines, and improving the discoverability of content across digital platforms.

B.Challenges in Caption Generation

Despite its pivotal role, image captioning confronts several formidable challenges, particularly in the generation of distinctive and contextually relevant captions. Standard datasets often furnish generic descriptions that fail to encapsulate the unique attributes of individual images, resulting in captions that are overly generic. Consequently, models trained on such datasets tend to produce similar captions for visually analogous images, curtailing their efficacy in real-world scenarios. Additionally, the conventional encoder-decoder framework utilized in image captioning may overlook local saliency in favor of global features. This necessitates the integration of attention mechanisms and semantic attention to enhance interpretability and elevate caption quality.

C.Evolution of Image Captioning Techniques

Evolution of Image Captioning Techniques

Over time, image captioning methodologies have undergone a remarkable evolution, driven by advancements in deep learning and the imperative to bridge the gap between visual perception and natural language understanding. Initially inspired by neural machine translation, modern approaches have transitioned from simplistic feature extraction to more nuanced models capable of discerning intricate semantic relationships within images.

Integration of Deep Learning Architectures

Contemporary image captioning techniques leverage sophisticated deep learning architectures, notably Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) models. CNNs serve as powerful encoders, extracting hierarchical features from images and encoding them into high-dimensional representations. These representations are then fed into LSTM decoders, which generate textual descriptions based on the encoded features. This encoder-decoder framework forms the backbone of many state-of-the-art image captioning systems, facilitating the seamless fusion of visual and linguistic information.

Emergence of Attention Mechanisms

A significant breakthrough in image captioning has been the integration of attention mechanisms, inspired by their success in machine translation tasks. Attention mechanisms enable models to selectively focus on salient regions within images while generating captions, thereby alleviating the limitations imposed by the traditional global feature extraction approach. By dynamically allocating attention to relevant image regions, these mechanisms enhance the richness and specificity of generated captions, leading to more accurate and contextually relevant descriptions.

Advancements in Semantic Understanding

Recent innovations in image captioning have also focused on enhancing semantic understanding through the incorporation of semantic attention mechanisms. Unlike traditional attention mechanisms that operate solely at the pixel level, semantic attention mechanisms enable models to discern semantic relationships between objects, attributes, and regions within images. By attending to semantically meaningful features, such as object categories and relationships, these mechanisms facilitate the generation of more coherent and semantically rich captions. Over time, image captioning methodologies have undergone a remarkable evolution,

driven by advancements in deep learning and the imperative to bridge the gap between visual perception and natural language understanding

V. ARCHITECTURE

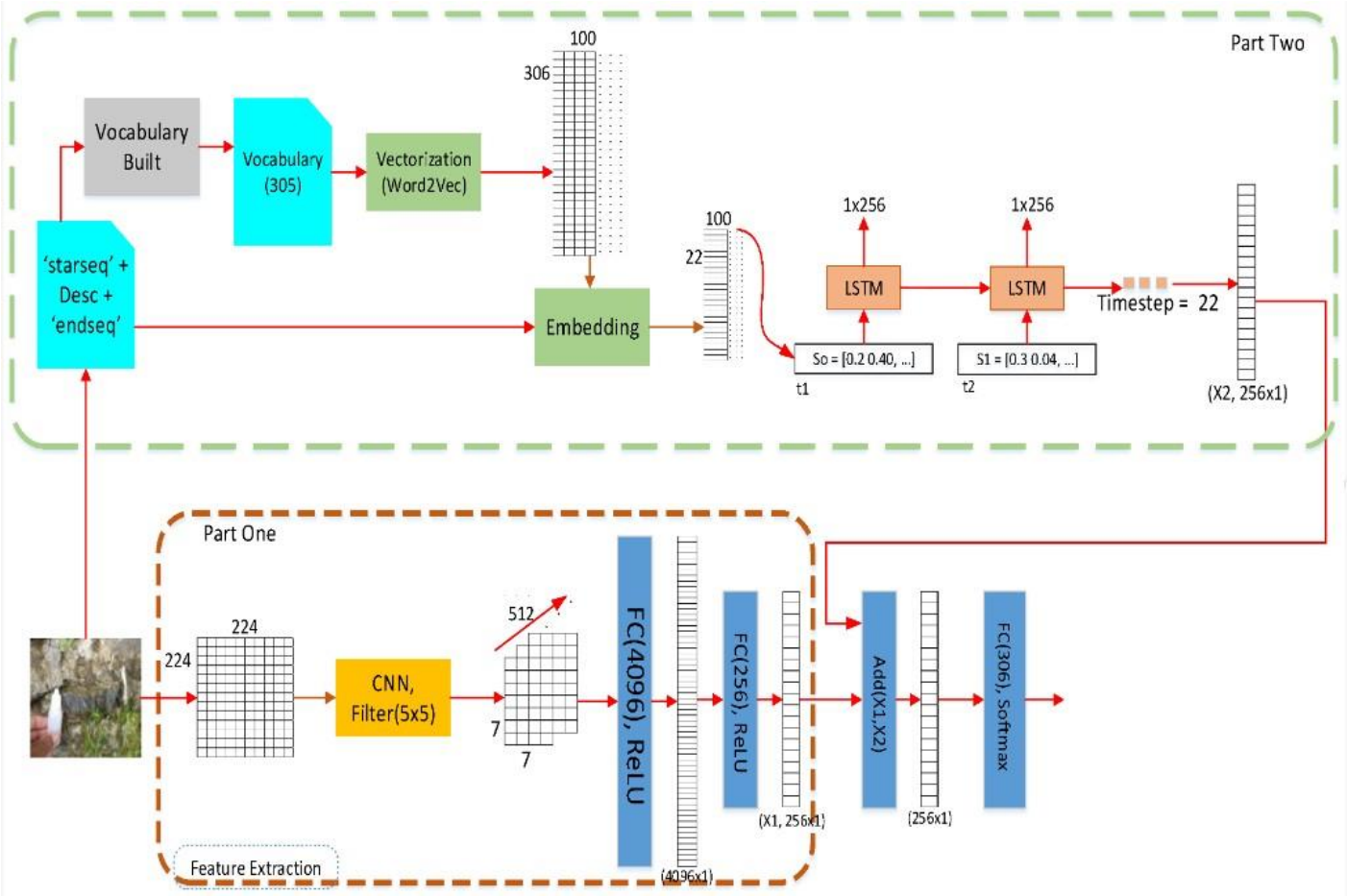

Figure 1, The general architecture diagram

As shown in Figure 1, The architecture delineated in the diagram embodies a sophisticated deep-learning framework meticulously engineered for the task of image captioning. Beginning with the initial phase termed "Feature Extraction," the model's journey commences with the input image traversing through a Convolutional Neural Network (CNN). This CNN serves as the vanguard, deploying a cascade of convolutional and pooling layers to systematically dissect the image into progressively abstract representations. At each layer, the CNN adeptly identifies discernible patterns, textures, and shapes, encapsulating them in feature maps. These feature maps, portrayed with dimensions of 224x224 and 512, serve as the bedrock of the subsequent stages, encapsulating salient visual cues across various scales and resolutions. Transitioning to the "Vocabulary Embedding" phase, the model embraces semantic enrichment by embedding the vocabulary words into dense vector representations. This semantic embedding process is facilitated by a dedicated Word Embedding Layer, which acts as a conduit for transforming raw textual data into meaningful numerical representations. Each word in the vocabulary is assigned a unique vector, with the dimensions of the vectors carefully calibrated to capture intricate semantic relationships between words. This embedding imbues the model with a nuanced understanding of the linguistic nuances inherent in the caption vocabulary, laying the groundwork for seamless integration of visual and semantic cues. As the journey progresses into "Combining Features and Vocabulary," the model orchestrates a harmonious fusion of visual and semantic information. The extracted visual features,

distilled through the CNN's discerning lens, are seamlessly amalgamated with the embedded vocabulary vectors. This fusion engenders a symbiotic relationship between visual and semantic modalities, endowing the model with a holistic understanding of the input image's content. By synthesizing disparate sources of information, the model achieves a comprehensive representation that transcends the sum of its parts, priming it for the arduous task of caption generation. Finally, the model embarks on the culminating phase of "Sequence Prediction," wherein the intricate tapestry of visual and semantic cues is woven into coherent captions. At the heart of this endeavor lies a Long Short-Term Memory (LSTM) network, renowned for its prowess in capturing temporal dependencies within sequential data. Armed with the amalgamated features and vocabulary vectors, the LSTM undertakes the formidable task of sequentially predicting the words that constitute the image caption. With each iteration, the LSTM navigates the labyrinthine landscape of linguistic intricacies, leveraging its memory cells to retain pertinent information and refine its predictions iteratively. In summation, the architecture epitomizes the synergy between cutting-edge deep learning techniques and the intricate nuances of image understanding and natural language processing. Through a meticulously orchestrated symphony of feature extraction, semantic embedding, and sequential prediction, the model transcends the realm of mere computational algorithms, emerging as a veritable maestro in the art of image captioning.

VI. ALGORITHM

1. Feature Extraction:

    Input: Image

    Initialize CNN with pre-trained weights

    Pass image through CNN to extract feature maps

    Output: Feature maps

    Vocabulary Embedding:

    Input: Vocabulary (list of words)

2. Initialize Word Embedding Layer

    Embed each word in the vocabulary into a dense vector representation

    Output: Embedded vocabulary vectors

    Combining Features and Vocabulary:

    Input: Feature maps, Embedded vocabulary vectors

3. Merge feature maps with embedded vocabulary vectors

    Output: Combined features and vocabulary

    Sequence Prediction:

    Input: Combined features and vocabulary

    Initialize the LSTM network for sequential processing

    Iterate over the combined input:

    Feed input sequentially into LSTM

4. LSTM Forecasts the subsequent word in the series of captions.
Repeat until the end-of-sequence token is generated or a maximum sequence length is reached

    Output: Predicted caption sequence

    Output Layer:

    Fully connected layer with ReLU activation (FC(4096), ReLU)

    Fully connected layer with softmax activation (FC(306), Softmax) to predict the next word

Training:

5. Initialize model parameters

    Define loss function (e.g., cross-entropy loss)

    Optimize model parameters using backpropagation and gradient descent

    Update weights iteratively to minimize loss on training data

    Repeat until convergence or a predefined number of epochs is reached

Inference:

6. Given a new image:

    Preprocess the image

    Extract features using the pre-trained CNN

Combine features with the embedded vocabulary vectors

Pass the combined input through the LSTM to generate the caption sequence

Use beam search or greedy decoding to generate the most likely caption

Output the generated caption

VII. MODULE DESCRIPTION

Module 1: Region Level Feature

This module is foundational for understanding the spatial relationships within an image, a crucial aspect for accurate caption generation. It begins by employing target-detection techniques to identify objects, their attributes, and their interconnections within the image. Once this information is obtained, a graph structure is constructed to represent these relationships visually. The graph structure provides a flexible and intuitive way to encode complex spatial dependencies among objects in the image. Graph convolutional neural networks (GCNs) are then utilized to extract features from the graph structure. GCNs are particularly suitable for this task because they can effectively capture the local and global relationships between nodes in a graph. By applying GCNs, the model can leverage the rich contextual information encoded in the graph to enhance the quality of feature representations. The extracted features are then used to guide the LSTM model in generating captions. By incorporating information about the spatial relationships between objects, the model can produce captions that are more contextually relevant and semantically accurate. This module plays a crucial role in bridging the gap between the visual content of the image and the textual description provided by the caption.

Module 2: Domain Object Pre-Filtering

In this module, the focus shifts to optimizing the input data fed into the captioning model by rearranging the order of image objects and region information. The process begins by utilizing a domain object dictionary, which is constructed during the inference of image captions. This dictionary contains predefined tags corresponding to specific objects or concepts relevant to the domain. The ordered pairs of image object tags and regional information are then sorted based on their priority according to the domain object dictionary. This prioritization ensures that objects of greater significance or relevance are given precedence in the captioning process. Additionally, objects that align with the domain object dictionary are copied and repeated to emphasize their importance in the generated captions.

By filtering and rearranging the input data in this manner, the model can focus its attention on the most relevant visual elements within the image, leading to more accurate and contextually appropriate captions. This module effectively enhances the interpretability and relevance of the generated captions, improving overall performance.

Module 3: Text Generation

Text generation lies at the heart of the captioning task, and this module delves into the intricacies of generating accurate and diverse descriptions of visual content. Traditional text generation methods often rely on deterministic approaches like greedy search or beam search, which may produce repetitive or monotonous captions. To overcome these limitations, stochastic methods such as top-k sampling are introduced. These methods allow the model to explore a wider range of possible captions by sampling from a distribution of likely words at each time step. By introducing randomness into the generation process, the model can produce more diverse and creative captions that better capture the nuances of the visual content. Furthermore, our module presents a semantically improved cross-modal fusion model intended to address modality disparity and inconsistent information. This approach efficiently aligns text and picture information inside a single semantic space by using multimodal representation learning methods, which makes caption creation more coherent and contextually appropriate.

Module 4: Visually-Aware Context Network

By improving the text's semantic representations, this module seeks to improve the calibre and applicability of produced captions. The module is composed of two main parts: a text encoder and a semantic improvement module. The text encoder maps the input text sequence into a semantic space, creating a preliminary semantic representation. Text improvement methods like rule enhancement and semantic upgrades are used to further improve phrases. A pooling layer, an atrous convolution, a residual network, and a multilayer perceptron (MLP) are among the components that are integrated by the

semantic improvement module. The extraction of higher-order characteristics and semantic information is facilitated by the multi-layer perceptual pyramid (MLP), which consists of many completely linked levels. Residual connection improves the expressiveness of feature representation and reduces problems such as gradient vanishing, which improves the quality of semantic representation. By incorporating dilatation, atrous convolution, also known as dilated convolution, increases the effective receptive field of the convolutional kernel, catching a greater variety of contextual information and enhancing text semantics understanding. Furthermore, a pooling layer lowers the spatial dimensionality of feature maps, improving processing speed. Through the synergy of these elements, the semantic improvement module efficiently collects and combines characteristics to provide a more complete and semantically enhanced representation of the input text. By improving the text's semantic content, this module helps to provide picture captions that are more accurate and relevant to the context.

Module 5: Evaluation Metrics

The evaluation of picture captioning algorithms' efficacy using a range of widely used metrics is the main objective of this module. Different viewpoints on caption quality are provided by these measures, which include BLEU-1 (B@1), BLEU-4 (B@4), METEOR (M), ROUGE-L (R-L), CIDEr, and SPICE. The accuracy of n-grams in produced captions is evaluated using BLEU-1 and BLEU-4 in comparison to reference captions. METEOR gives a fair assessment of memory and accuracy while taking synonymy, stemming, and paraphrasing into account. The overlap between the produced and reference captions' longest common subsequence is measured using ROUGE-L. CIDEr measures the agreement between produced descriptions and human captions, whereas SPICE measures the propositional semantic content of captions. Researchers may evaluate the effectiveness of captioning models in-depth and pinpoint areas for development by using this broad range of assessment measures. The emphasis of this session is on the value of thorough assessment techniques in raising the bar for picture captioned.

VIII. RESULT

As shown in Figure 2, The results provide a detailed insight into the performance of the machine learning model, particularly focusing on training and validation accuracy curves. The training accuracy curve depicts how effectively the model classifies the training data over epochs, starting at approximately 0.825 and progressively improving to nearly 1.000 by epoch 9. This upward trend signifies that the model effectively learns from the training data and becomes increasingly adept at classifying images as training progresses. Conversely, the validation accuracy curve illustrates the model's performance on unseen data, starting around 0.875 and reaching approximately 0.925 by epoch 9. Although there is improvement over time, the validation accuracy consistently lags behind the training accuracy. This discrepancy suggests potential overfitting, where the model excels at capturing intricate patterns and noise within the training data but struggles to generalize to new, unseen data. Overfitting poses a significant challenge in machine learning, as it compromises the model's ability to generalize beyond the training data. To address this issue, several strategies can be employed. One approach involves reducing the complexity of the model by decreasing the number of layers or units, thereby mitigating its tendency to memorize specific details of the training data. Additionally, regularization techniques can be implemented to penalize the model for complexity, encouraging it to focus on learning more generalizable patterns. Additionally, data augmentation methods provide a workable way to fictitiously increase the training dataset's size and variety. Data augmentation enhances the model's exposure to data changes by applying random alterations, such as flips, cuts, or color tweaks, to pre-existing pictures. This promotes generalization and resilience in the model. Essentially, the validation accuracy curve emphasizes the need to minimize possible overfitting, while the training accuracy curve highlights the model's ability to learn from the training data. Model simplification, regularisation, and data augmentation are some of the tactics that may be used to improve overall performance by strengthening the model's ability to generalize to new data.

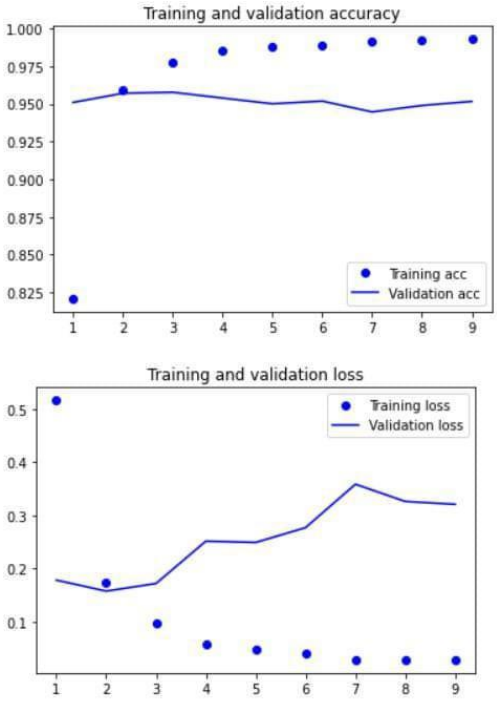

Figure 2, the output of the algorithm.

IX. CONCLUSION

This paper introduces a novel approach to object-controllable image captioning, addressing a key limitation in existing methods by leveraging prior information on detected objects and their relationships. Two key parts make up the suggested method, known as the information-augmented graph encoder: a multi-relational weighted graph encoder and an information-augmented embedding module. The control signal, node attributes, and previous data are fused using a dynamic attention model to improve the captioning process. In addition, a similarity loss mechanism is developed to promote the creation of varied captions. The suggested approach offers state-of-the-art performance in controlled picture captioning, as shown by extensive testing on the Flickr30k Entities dataset. The study also encompasses a comprehensive review of datasets and evaluation metrics commonly used in training and evaluating image captioning systems. Moreover, it delves into the concept of pre-training, highlighting the emergence of vision-language pre-training as a valuable approach in computer vision tasks. By examining various captioning models, including those based on the vanilla transformer architecture and those incorporating vision-language pre-training, the paper conducts a detailed analysis of design choices and performance comparisons among these models. Overall, our research underscores the effectiveness of neural network techniques in addressing image-captioning tasks. Through rigorous experimentation, we demonstrate that the proposed methods contribute to reducing model size without significant quality loss. This study not only advances the field of object-controllable image captioning but also sheds light on the potential of neural network approaches in tackling complex computer vision challenges.

X.REFERENCES

[1], Sio-Kei Im, Ka-Hou Chan Context-Adaptive-Based Image Captioning by Bi-CARU IEEE Access, 2023.

[2], Yong Wang, Wenkai Zhang, Zhengyuan Zhang, Xin Gao, Xian Sun Multiscale Multi interaction Network for Remote Sensing Image Captioning IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2022.

[3], Taku Suzuki, Daisuke Sato, Yoshihiro Sugaya, Tomo Miyazaki, Shinichiro Omachi Important Region Estimation Using Image Captioning IEEE Access, 2022.

[4], Hyeryun Park, Kyungmo Kim, Seongkeun Park, Jinwook Choi Medical Image Captioning Model to Convey More Details: Methodological Comparison of Feature Difference Generation IEEE Access, 2021.

[5], Zhenyu Yang, Qiao Liu ATT-BM-SOM: A Framework of Effectively Choosing Image Information and Optimizing Syntax for Image Captioning IEEE Access, 2020.

[6], Eric Ke Wang, Xun Zhang, Fan Wang, Tsu-Yang Wu, Chien-Ming Chen Multilayer Dense Attention Model for Image Caption IEEE Access, 2019.

[7], Khang Nguyen, Doanh C. Bui, Truc Trinh, Nguyen D. Vo EAES: Effective Augmented Embedding Spaces for Text-Based Image Captioning IEEE Access, 2022.

[8], Shiwei Wang, Long Lan, Xiang Zhang, Guohua Dong, Zhigang Luo Cascade Semantic Fusion for Image Captioning IEEE Access, 2019

[9], Zhou Lei, Congcong Zhou, Shengbo Chen, Yiyong Huang, Xianrui Liu A Sparse Transformer-Based Approach for Image Captioning IEEE Access, 2020.

[10], Suya Zhang, Yana Zhang, Zeyu Chen, Zhaohui Li VSAM-Based Visual Keyword Generation for Image Caption IEEE Access, 2021.

[11], Jian Wang, Jie Feng Hybrid Attention Distribution and

Factorized Embedding Matrix in Image Captioning IEEE Access, 2020.

[12], 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.

[13], T. Wolf et al., Transformers: State-of-the-art natural language processing, Proc. Conf. Empirical Methods Natural Lang. Process. Syst. Demonstrations, pp. 38-45, 2020.

[14], 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.

[15], 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.

[16], 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.

[17], B. Qu, X. Li, D. Tao and X. Lu, Deep semantic understanding of high-resolution remote sensing image, Proc. Int. Conf. Comput., pp. 1-5, 2016.

[18], X. Lu, B. Wang, X. Zheng, and X. Li, Exploring models and data for remote sensing image caption generation in IEEE Trans. Geosci. Remote Sens., vol. 56, no. 4, pp. 2183-2195, Apr. 2018.

[19], Z. Zhang, W. Zhang, M. Yan, X. Gao, K. Fu and X. Sun, Global visual feature and linguistic state guided attention for remote sensing image captioning, IEEE Trans. Geosci. Remote Sens., pp. 313-318, 2021.

[20], C. Zhang, G. Li, and S. Du, Multi-scale dense networks for hyperspectral remote sensing image classification, IEEE Trans. Geosci. Remote Sens., vol. 57, no. 11, pp. 9201-9222, Nov. 2019.

[21], Z. Zhang, H. Wang, F. Xu, and Y.-Q. Jin, Complex-valued convolutional neural network and its application in polarimetric SAR image classification, IEEE Trans. Geosci. Remote Sens., vol. 55, no. 12, pp. 7177-7188, Dec. 2017.

[22], C. Zhang, T. Zou, and Z. Wang, A fast target detection algorithm for high-resolution sar imagery in J. Remote Sens, vol. 9, no. 1, pp. 45-49, 2005.

[23], W. Siyu, G. Xin, S. Hao, Z. Xinwei and S. Xian, An aircraft detection method based on convolutional neural networks in high-resolution sar images, J. Radars, vol. 6, no. 2, pp. 195-203, 2017.

[24], Z. Zheng, Y. Zhong, J. Wang, and A. Ma, Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery, Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pp. 4096-4105, Jun. 2020.

[25], D. Marmanis, J. D. Wegner, S. Galliani, K. Schindler, M. Datcu and U. Stilla, Semantic segmentation of aerial images with an ensemble of class, ISPRS Annals Photogramm. Remote Sens. Spatial Inf. Sci., vol. 3, pp. 473-480, 2016.

[26], O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, Show and tell: A neural image caption generator, Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 3156-3164, 2015.

[27], D. Reinsel, J. Gantz and J. Rydning, The digitization of the world from edge to core, Needham, MA, USA, 2018. » Cisco Visual Networking Index: Forecast and Trends 2017â€“2022 White Paper, San Jose, CA, USA, 2019.

[28], S. Shioiri, Y. Sato, Y. Horaguchi, H. Muraoka and M. Nihei, Quali-informatics in the society with yotta scale data, Proc. IEEE Int. Symp. Circuits Syst. (ISCAS), pp. 1-4, May 2021.

[29], V. Smil, Data world: Racing toward yotta, IEEE Spectr., vol. 56, no. 7, pp. 20, Jul. 2019.

[30], O. Issa and T. Shanableh, CNN and HEVC video coding features for static video summarization, IEEE Access, vol. 10, pp. 72080-72091, 2022.

[31], A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, et al., Every picture tells a story: Generating sentences from images, Proc. Eur. Conf. Comput. Vis, pp. 15-29, 2010.

[32], M. Hodosh, P. Young and J. Hockenmaier, Framing image description as a ranking task: Data models and evaluation metrics, J. Artif. Intell. Res., vol. 47, pp. 853-899, Aug. 2013.

[33], R. Mason and E. Charniak, Nonparametric method for data-driven image captioning, Proc. 52nd Annu. Meeting Assoc. for Comput. Linguistics (Short Papers), vol. 2, pp. 592-598, 2014.

[34], A. Gupta, Y. Verma and C. V. Jawahar, Choosing linguistics over vision to describe images, Proc. 26th AAAI Conf. Artif. Intell., pp. 606-612, 2012.

[35], P. Kuznetsova, V. Ordonez, T. L. Berg, and Y. Choi, TreeTalk: Composition and compression of trees for image descriptions, Trans. Assoc. Comput. Linguistics, vol. 2, pp. 351-362, Dec. 2014.

[36], L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, et al., SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 5659- 5667, Jul. 2017.

[37], J. Lu, C. Xiong, D. Parikh and R. Socher, Knowing when to look: Adaptive attention via a visual sentinel for image captioning, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 375-383, Jul. 2017.

[38], S. Ren, K. He, R. Girshick and J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, Proc. Adv. Neural Inf. Process. Syst., pp. 91-99, 2015.

[39], Z. Zhu, Z. Xue and Z. Yuan, Topic-guided attention for image captioning, Proc. 25th IEEE Int. Conf. Image Process. (ICIP), pp. 2615-2619, Oct. 2018.

[40], Q. You, H. Jin, Z. Wang, C. Fang and J. Luo, Image captioning

with semantic attention, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 4651-4659, Jun. 2016.

[41], J. C.-W. Lin, Y. Zhang, B. Zhang, P. Fournier-Viger and Y. Djenouri, Hiding sensitive itemsets with multiple objective optimization, Soft Comput., pp. 1-19, Feb. 2019.

[42], J.-S. Pan, L. Kong, T.-W. Sung, P.-W. Tsai and V. SnÃ¡Å¡el, \$alpha\$ -fraction first strategy for the hierarchical model in wireless sensor networks, J. Internet Technol., vol. 19, pp. 1717-1726, 2018.

[43], J. Guan and E. Wang, Repeated review based image captioning for image evidence review, Signal Process. Image Commun., vol. 63, pp. 141-148, Apr. 2018.

[44], O. Vinyals, A. Toshev, S. Bengio and D. Erhan, Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge, IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 4, pp. 652-663, Apr. 2017.

[45], A. Karpathy and L. Fei-Fei, Deep visual-semantic alignments for generating image descriptions, Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 3128-3137, Jun. 2015.

[46], K. Xu et al., Show attend and tell: Neural image caption generation with visual attention, Comput. Sci., vol. 2015, pp. 2048-2057, Feb. 2015.

[47], P. Anderson, Bottom-up and top-down attention for image captioning and visual question answering, Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pp. 6077-6086, Jun. 2018.

[48], N. Li and Z. Chen, Image captioning with visual-semantic LSTM, Proc. IJCAI, pp. 793-799, 2018.

[49], H. Fang et al., From captions to visual concepts and back, Proc. CVPR, pp. 1473-1482, Jun. 2015.

[50], G. Kulkarni et al., Baby talk: Understanding and generating simple image descriptions, Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 1601-1608, Jun. 2011.

[51], O. Sidorov, R. Hu, M. Rohrbach and A. Singh, TextCaps: A dataset for image captioning with reading comprehension, Proc. Eur. Conf. Comput. Vis., pp. 742-758, 2020.

[52], O. Sidorov, R. Hu, M. Rohrbach and A. Singh, TextCaps: A dataset for image captioning with reading comprehension, Proc. Eur. Conf. Comput. Vis., pp. 742-758, 2020.