IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Comprehensive Survey on Image Description Generation Techniques

[1]Praveen Kumar Tripathi, [2]Dr. Shambhu Bharadwaj

[1]Research Scholar, [2]Associate Professor

[1,2] Department of Computer Science & Engineering

[1,2]College of Computing Sciences & Information Technology, TMU, Moradabad

Abstract

Image captioning systems combine computer vision and natural language processing to generate accurate descriptions of images. The computer vision module extracts relevant information from images, while the NLP module constructs coherent and meaningful captions. This technology has gained significant interest due to its applications in analyzing large image datasets, detecting patterns for machine learning, and developing assistive software for the visually impaired. This study focuses on image captioning using neural networks, specifically ResNet as the encoder for image data, CNN for video subtitles, and LSTM (Long-Short-Term Memory) as the decoder.

Keywords - Deep learning, Natural Language Processing, Image Captioning, ResNet, LSTM, Recurrent Neural Network, Convolutional Neural Network.

1. Introduction

Image captioning is the automated process of generating descriptive text for images, leveraging computer vision and natural language processing. This technology involves object detection, relationship analysis, and attributes prediction, utilizing the unique qualities of each object. Deep neural networks, particularly CNNs and RNNs with LSTM, demonstrate promising capabilities in object detection and learning from previous samples. By combining CNNs for visual identification and RNNs for language modelling, the ResNet-LSTM model can be employed to develop an image caption generator. This system utilizes convolutional matrices to extract image features, which are then transformed into feature vectors and used by LSTM networks to generate descriptive captions [1]. The integration of CNNs and LSTMs enables the creation of a comprehensive image caption generator.

2. Literature Review

A comprehensive literature review on image description generation techniques would involve an in-depth analysis of the various methods and approaches used in the field of computer vision and natural language processing to automatically generate textual descriptions of images. In previous research, the primary focus was on recognizing objects in images based on predefined categories. However, this method is limiting compared to the way individuals naturally converse. Recent studies have significantly progressed in the identification, categorization, and naming of objects in images. However, achieving a detailed description of a complex scene necessitates a more profound comprehension. This involves understanding the unfolding events in the scene, discerning the relationships among different objects, and expressing it in a natural-sounding manner.

Moses et al. [4] developed a generative CNN-LSTM model that surpasses human baselines by a significant margin of 2.7 BLEU-4 points, closely approaching state-of-the-art performance with a difference of merely 3.8 CIDEr points. Experiments on the MS COCO dataset demonstrate the model's capability to generate accurate and coherent captions. Hyper parameter tuning, including dropout and the number of LSTM layers,

effectively mitigates overfitting. Furthermore, the study reveals that semantically similar words, such as "plate" and "bowl", exert a similar influence on the LSTM hidden state, whereas semantically distant words, like "vase" and "food", cause a divergence in the hidden state. This highlights the semantic significance of the interplay between learned word embeddings and LSTM hidden states, showcasing the model's ability to capture nuanced relationships between words [2].

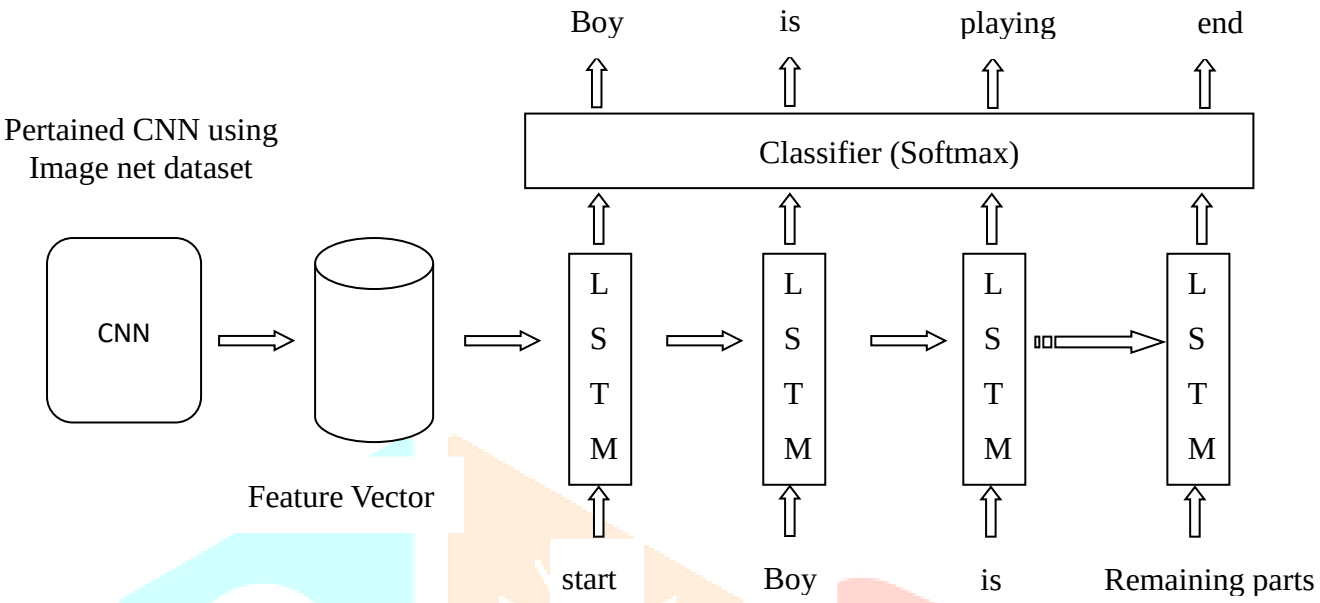

Fig. 1 CNN – LSTM model

R. Subash et al. [12] showed that previous methods, which directly compared the query image with database images, were limited to generating generic and predefined natural language descriptions. However, these approaches restricted the diversity of generated captions, relying on closed visual vocabularies that are inadequate compared to the vast range of images humans can create. The proposed model, on the other hand, learns to construct sentences from scratch using the training data, without assuming predetermined formats, standards, or classifications. By leveraging convolutional neural networks to extract essential image components and natural language processing techniques, the model generates logical statements and captions through a probabilistic approach, enabling more flexible and diverse image descriptions.

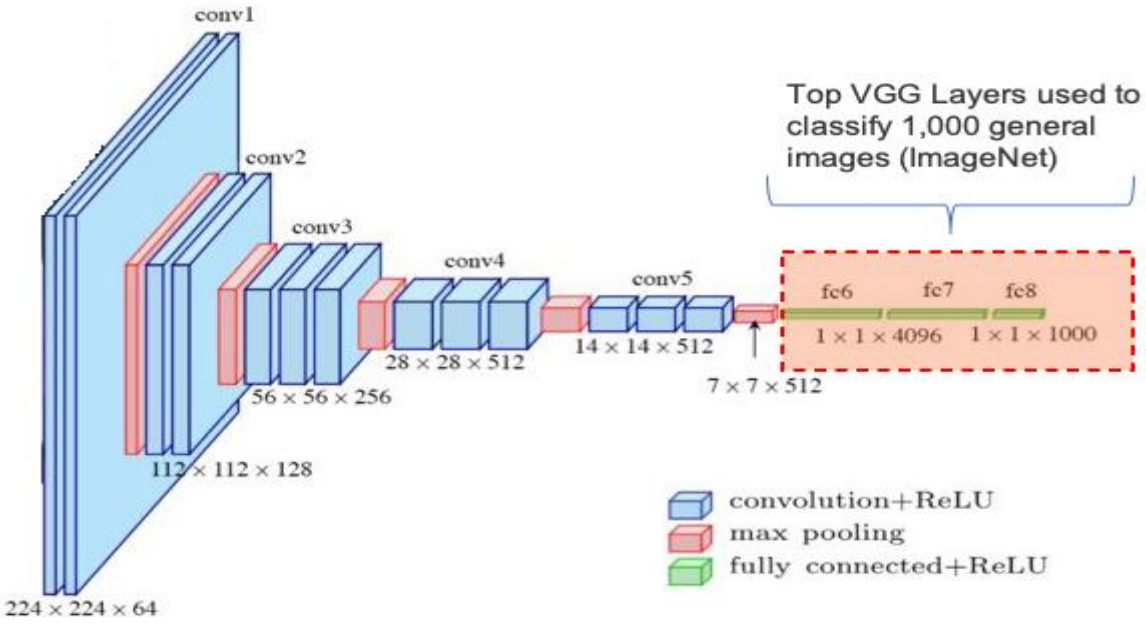

Fig. 2 VGG 16 Architecture

Abisha Anto Ignatious. L et al. [13] proposed an enhanced semantic-driven CNN-LSTM model, incorporating feature extraction, semantic keyword retrieval, face detection, and encoder-decoder LSTM networks. A pre-trained CNN extracts image features, while a semantic keyword extraction module identifies objects and assigns names using semantic tags. This approach enhances caption descriptive

capabilities. LSTM-based word embedding generates each word of the captions. Additionally, a face recognition system utilizing a faces dataset of 232 celebrities detects and recognizes famous faces, replacing instances with the person's name to create personalized subtitles. The accuracy of generated captions is evaluated using BLEU and METEOR scores, ensuring a comprehensive assessment of the model's performance.

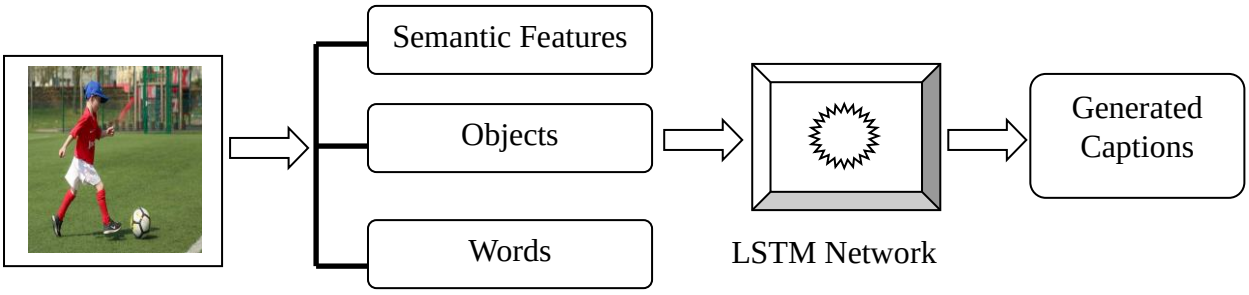

Fig. 3 Architectural Language model

Himanshu Sharma et al. [14] proposed a novel approach that leverages both visual and additional information from knowledge sources like ConceptNet to initialize images. They demonstrated the efficacy of their method using publicly available datasets, Flickr8k and Flickr30k. The results show that their proposed model outperforms state-of-the-art techniques in image caption generation, achieving improved performance. This work paves the way for future advancements in image captioning, which will be discussed in the subsequent section.

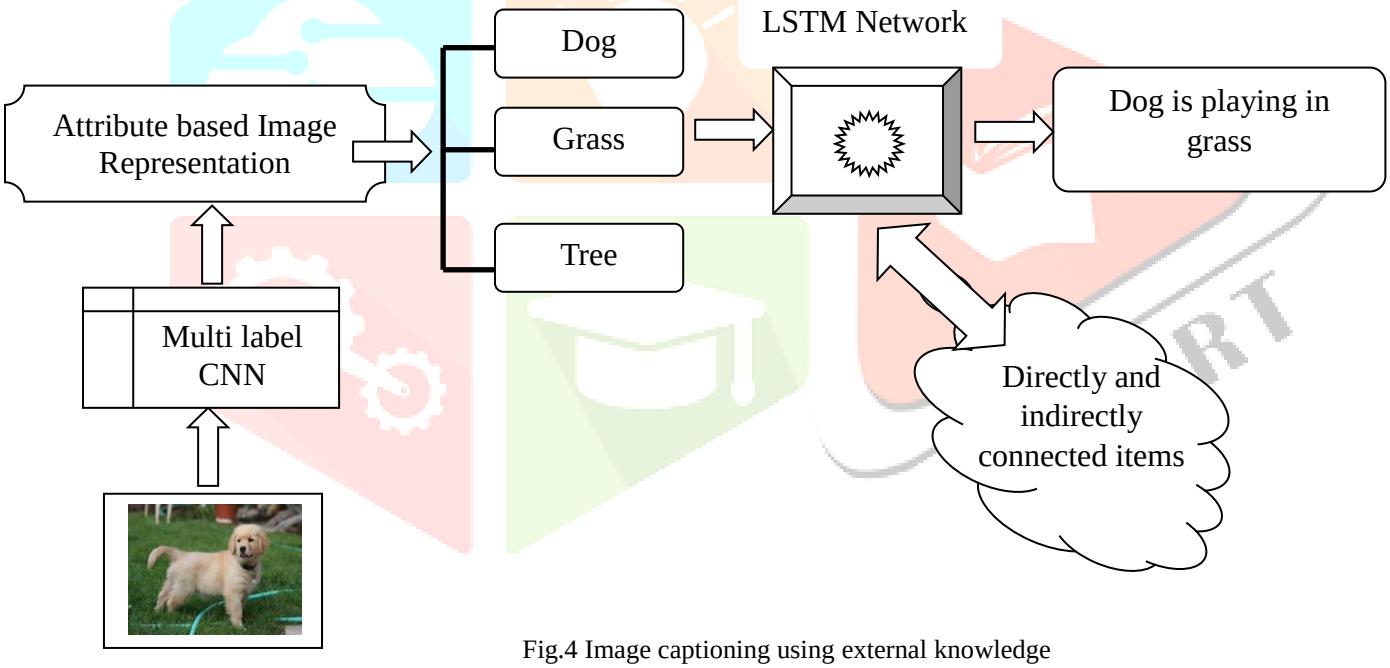

Fig.4 Image captioning using external knowledge

G. Geetha et al. [15] developed an algorithm aimed at enhancing global understanding of deforestation dynamics, encompassing its spatial distribution, drivers, and consequences. The advent of advanced satellite imaging technologies is poised to facilitate more precise monitoring and analysis of environmental changes, including deforestation. The Amazon rainforest, for instance, has suffered significant losses, with approximately 20% of its area cleared over the past four decades. To estimate and analyze forest cover, the application leverages multiclass multi-label image classification frameworks, including gate recurrent unit label captioning and sparse cross-entropy. Deep convolutional neural networks (CNNs) are trained on satellite imagery to extract visual features, which are then integrated into a hybrid architecture combining a pre-trained VGG-19 encoder and a GRU decoder trained on ImageNet data.

Wang et al. [3, 17, 19] introduced the Bidirectional Long Short-Term Memory (Bi-LSTM) architecture, aiming to improve image captioning by leveraging both past and future context in sequence-to-sequence learning. By integrating Convolutional Neural Networks (CNNs) for visual feature extraction and deep Bi-LSTM for text translation, their model demonstrated enhanced performance in generating extended word

sequences. The Bi-LSTM approach showed improved capability in producing successive words, leading to higher BLEU scores and substantiating the effectiveness of this architecture in image captioning tasks.

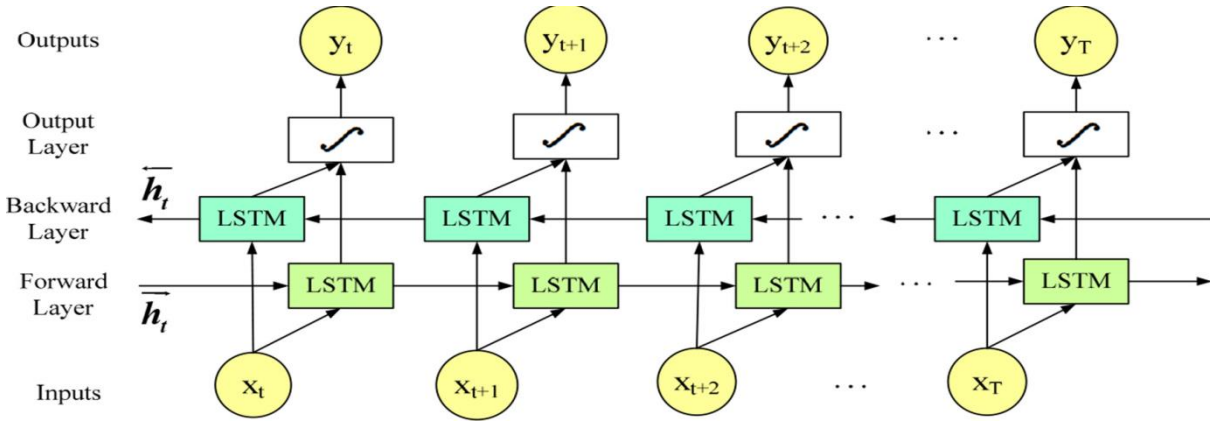

Fig. 5 Multimodel Bidirectional LSTM

Figure 5 illustrates the multi-layer architecture of the proposed model, comprising the sentence embedding layer (L1), Temporal LSTM layer (L2), Multimodal LSTM layer (L3i), and Softmax layer. This model generates sentences word-by-word over time by processing input sentences in both forward (blue arrow) and backward (red arrow) directions. The model is end-to-end trainable by minimizing a joint loss function.

Bin Yi et al. [8] introduced a novel video captioning framework that integrates soft attention and bidirectional long-short-term memory (BiLSTM) to enhance global video representations and recognize persistent emotions. This framework aims to improve video captioning by leveraging the strengths of both soft attention and BiLSTM in capturing temporal dependencies and contextual information.

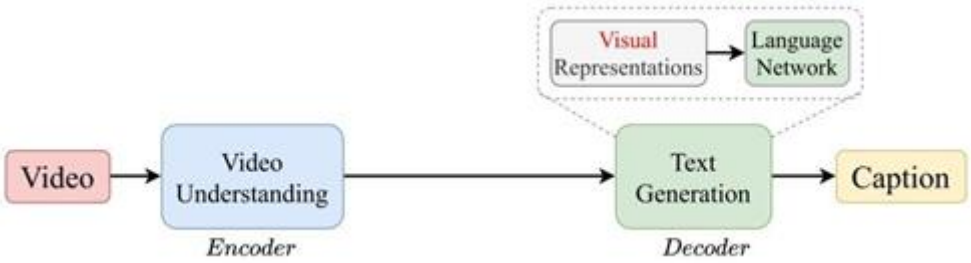

Fig. 6 Vedio Captioning model

The visual encoder and natural-language generator components were implemented using two LSTM networks, providing a foundation for the model's architecture. Building on this, the initial CNN features were leveraged to capture the bidirectional global temporal structure in video clips, and a bidirectional LSTM network was employed to encode video sequences. The language decoder then generated text from the encoded video sentences. The proposed model demonstrated superior performance compared to several state-of-the-art approaches on the MSVD and MSR-VTT 10K datasets, showcasing its effectiveness in video captioning tasks.

Md Zakir et al. [9] proposed the Bi-Directional Self-Attention (Bi-SAN) method for image captioning, leveraging the self-attention mechanism that has gained popularity in various sequence modeling tasks. This attention-based approach dispenses with the need for LSTM/CNN, instead employing bi-directional self-attention to caption images by shifting focus both backward and forward. By computing attention in both directions using front and backward masks, Bi-SAN enables parallel computation, overcoming the temporal dependence limitation of LSTMs. This approach yields rich feature representations through bi-SAN and inter-attention, without requiring recurrence or convolution. The proposed Bi-SAN-based image captioning methodology demonstrates superior performance on ROUGH-L metrics and BLEU-1, 3, 4, outperforming traditional methods.

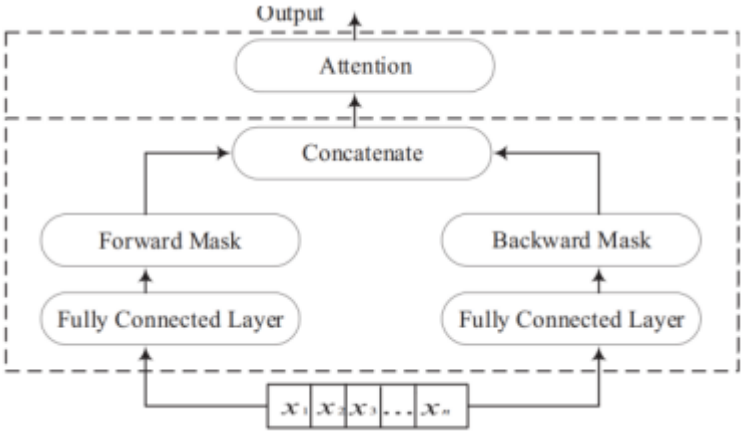

Fig. 7 Bidirectional Self attention sequence modelling

Figure 7 illustrates the bi-directional self-attention model employed in the proposed sequence modelling approach [8].

Xinxin et al. [10] introduced a novel image captioning task that requires reasoning about both past and future word order to generate relevant descriptions. They demonstrated the effectiveness of their approach on the Visual Madlibs dataset and consistently outperformed baseline methods. They also proposed the first approximation inference approach for bidirectional neural sequence models' 1-Best (and M-Best) decoding, utilizing knowledge from both past and future to "fill in the gaps."

Stefan Lee et al. [7] developed a stack parallel LSTM model with triple attention for image captioning. The model consists of three stages, where the attention model is used as input, combined with LSTM, and used as output, respectively. Evaluating the model on the MS-COCO dataset showed that the triple attention (TA-LSTM) approach, which incorporates visually relevant information at every stage, enhances performance compared to traditional LSTM.

Ghadah Alabduljabbar et al. [16] proposed a model designed to perform well with minimal engineering effort. Unlike approaches that focus solely on visual information, their model utilizes fine-grained visual data obtained using ResNet101 and a refining model to generate a clean representation of retrieved object features.

Yuqing Peng et al. [17] addressed the limitations of existing image semantic comprehension models, which often misinterpret or overlook scene identification, resulting in low accuracy rates for descriptive sentences. They proposed a comprehensive image semantic comprehension model that combines scene components, leveraging the text volume of the LDA analysis corpus to determine the theme. The model extracts global image features using ResNet and deep scene features using Places365-CNNs, incorporating scene data from the corpus and images.

Xin Jin et al. [22] introduced Aesthetic Image Captioning (AIC), a specialized end-to-end network that utilizes images and associated aesthetic judgments for training, generating textual interpretations of visual aesthetics. The model employs a sequence generation approach, typically using ResNet101 + LSTM, where the encoder permanently encodes the data, and the decoder generates the sequence word-by-word.

Xianrui Li et al. [14] leveraged the ImageNet classification challenge to pre-train the ResNet-101 model, which was then used as the foundation for the AttriNET attribute detection model. The transfer learning framework was built upon the pre-trained ResNet-101, enabling the integration of attribute detection and visual attention. This approach facilitated the development of a unified framework for simultaneously detecting attributes and generating visual attention maps, enhancing the model's capability to capture fine-grained image features.

Shobiya L, Pradheesha R, and Prof. Kala [19] proposed a generative model that integrates advances in computer vision and machine translation to generate natural sentences describing an image. The model's architecture comprises Residual Networks (ResNet), RNN, and LSTM-based sentence generation. Evaluation metrics demonstrate superior performance compared to previous benchmark models.

Aishwarya Maroju et al. [21] presented a technique employing LSTM for decoding and ResNet architecture for encoding, trained on the Flickr8k dataset. The ResNet-LSTM model, executed on a Graphic Processing Unit, exhibits improved accuracy compared to the CNN-RNN and VGG models, showcasing the effectiveness of this approach in image captioning tasks.

Yan Chu et al. [19] introduced a joint model, AICRL, which performs automatic image captioning using soft attention-based ResNet50. The AICRL model consists of an encoder and decoder, with ResNet50 enhancing overall efficiency. The soft attention technique improves performance across all metrics, including METEOR, CIDER, and BLEU-4, demonstrating the effectiveness of this approach in image captioning tasks.

Shruti Mundargi et al. [20] employed the Resnet-LSTM model for image captioning, utilizing LSTM for decoding and Resnet for encoding. The model's efficiency was enhanced by executing it on a graphics processing unit, with Resnet50 reducing computational costs and training time.

Janvi Jambhale et al. [19] conducted an analysis of the visual attention mechanism, exploring captions generated using three preprocessing models: Inception V3, VGG19, and Resnet50. The crucial components of images were identified, yielding efficient results through the Ensemble learning technique.

Abhishek Sethi et al. [20] presented a study emphasizing the impact of attention-based learning on automatically generated Hindi captions, utilizing the Hindi-language Flickr8K dataset to evaluate performance in terms of BLEU score. The methodology employed a parallel architecture with two pipelines: the first pipeline generated image features using pre-trained CNN models (Resnet50 and VGG16), while the second pipeline processed reference captions using LSTM. The combined output from both pipelines generated image captions through a fusion layer composed of 256 neurons.

3. DISCUSSION

3.1 Computational Resources and Model Architecture

The encoder-decoder framework, a versatile and powerful paradigm, is widely employed in image captioning models. This architecture is often referred to as a CNN+RNN structure, where Convolutional Neural Networks (CNNs) serve as the encoder and Recurrent Neural Networks (RNNs) as the decoder. The encoder extracts high-level features from the input image, acting as a "reader," while the decoder generates words to form a comprehensive and grammatically correct sentence.

3.1.1 CNN-based Image Encoding

Deep-learning image captioning models rely on CNNs for image encoding due to their ability to detect patterns in pixel intensity values. Various CNN architectures are employed in the Image Encoding module, including AlexNet, VGGNet, GoogLeNet/Inception-V1, ResNet, Inception V3, and Inception V1. VGGNet is a popular choice for feature extraction due to its simplicity and efficacy, while ResNet is equally employed as an encoder, offering superior computational efficiency.

3.1.2 LSTM-based Decoding

Long Short-Term Memory (LSTM) networks, a type of Recurrent Neural Networks (RNNs), are capable of capturing long-range dependencies, making them suitable for challenging tasks such as speech recognition, machine translation, and more. Conventional RNNs often suffer from vanishing gradients during training, leading to poor performance. In contrast, LSTM networks are designed to handle delays of varying lengths between significant events in a time series, enabling effective classification, analysis, and prediction. LSTM outperforms traditional RNNs in performance, overcoming the limitations of short-term memory. By processing inputs while selectively retaining relevant information and discarding irrelevant data, LSTMs demonstrate superior efficacy.

3.2 Datasets

1. Flickr30k, a widely-used dataset for image captioning, consists of 31,783 images. The dataset is divided into 29,783 images for training, 1,000 images for testing, and 1,000 images for validation. Notably, each image in this collection is accompanied by five distinct captions.

2. MS COCO is another prominent dataset for image captioning, comprising 123,287 images. The dataset is partitioned into 113,287 images for training, 5,000 images for validation, and 5,000 images for testing. Each image in this dataset is paired with five human-generated captions, providing a rich resource for image captioning research.

Comparisons of reference captions on datasets MS COCO, Flickr8K, and Flickr30K

Datasets	Vocab Size	Max Length	Total Words	Top-10 Words with Higher Occurrences
MS COCO	9486	49	6,421,733	a, on, of, the, in, with, and, is, man, to
Flickr8K	2629	37	422,800	a, in, the, on, is, and, dog, with, man, of
Flickr30K	7648	78	1,892,755	a, in, the, on, and, man, is, of, with, woman

Table 1 Analysis of datasets

4. IV. COMPARATIVE STUDY

4.1 Results and conclusion of t h e existing system

The research by Soh, Moses, and colleagues [1] presents findings on the 2D projection of sentence hidden states. The first sentence pair reveals that words with similar semantic meanings move the hidden state in the same direction, despite being conditioned on different image vectors. The second sentence pair shows that the emitted sequence similarly shifts the hidden dimension, with the sentence representation in the hidden state only diverging when the sentence starts to describe image variations.

According to R. Subash et al. [2], the Bilingual Evaluation Understudy (BLEU) score assesses a computer's understanding of content. As a key metric, BLEU strongly correlates with human evaluation. The researchers compared their model's performance on the MSCOCO dataset to existing models and found that it outperforms them, demonstrating its effectiveness.

Abisha Anto Ignatious. L et al. [3] implemented the proposed model using Keras with a Tensorflow backend, utilizing an LSTM language model with 128 cells. The dataset was split into 60% for training, 20% for validation, and 20% for testing.

Himanshu Sharma et al. [4] found that the proposed model surpasses state-of-the-art methods in image description generation. They evaluated the model's performance on the widely used, publicly available Flickr8k and Flickr30k datasets, demonstrating its effectiveness.

G. Geetha et al. [5] observed that the Attention-based technique's performance varies depending on the convolutional neural network (CNN) used, with different CNN architectures leveraging convolutional features in attention-based approaches. Experimental results demonstrate the impact of different CNNs on the technique's outcomes.

Wang et al. [6] found that the Bi-LSTM architecture excelled in generating consecutive words, achieving higher BLEU scores and longer word sequences. The results confirmed that the proposed bidirectional LSTM method optimized image captions. Furthermore, the bidirectional LSTM model demonstrated the benefits of multi-task learning in enhancing model performance.

Bin Yi et al. [7] conducted experiments on the MSR-VTT 10K and MSVD datasets, outperforming several state-of-the-art methods. Their approach utilized a BiLSTM and soft attention mechanism to generate better global representations of videos, enhancing the detection of long-lasting motions.

Hossain, Md Zakir et al. [8] employed both forward and backward attention, achieving performance comparable to cutting-edge techniques. Their approach demonstrated that bi-SAN can effectively capture rich feature representations without relying on recurrence or convolution, resulting in reduced computational time.

Zhu, Xinxin et al. [9] observed improved performance compared to base models, achieving test results with a single model rather than relying on ensemble processes. Experimental results showed that the triple attention mechanism and LS-LSTM can generate more readable and fluent sentences.

Qing et al. [10] proposed a novel Bidirectional Beam Search (BiBS) approach, addressing the limitation of unidirectional RNNs that generate illogical outputs due to their inability to consider both previous and upcoming contexts. This marks the first top-B MAP inference approach for Bidirectional RNNs, enabling more effective processing.

Ghadah Alabduljabbar et al. [11] introduced a new approach that leverages extracted feature characteristics and an attention-on-attention mechanism to describe interactions between image elements, moving beyond simply enhancing visual features.

Yuqing Peng et al. [12] demonstrated that their double Long short-term memory fusion model and scene factors prioritize the impact of scene on overall semantic meaning, outperforming reference models.

Xin Jin et al. [13] constructed a new image aesthetics dataset and developed a latent Dirichlet assessment model to filter it, resulting in the FAE-Captions dataset. They also proposed a novel convolutional neural network model that generates optimal aesthetic comments, showcasing the effectiveness of their approach and dataset.

Xianrui Li et al. [14] developed an effective image captioning approach for clothing images, utilizing a joint attribute detection and visual attention framework to generate captions that capture the image's characteristics.

Shobiya L et al. [15] proposed a single encoder-decoder architecture, employing ResNet101 as the encoder to extract visual features from the image and an LSTM language model as the decoder to generate descriptive sentences. Additionally, they integrated a sensitive attention model with the LSTM to enable targeted learning on specific image regions, enhancing performance.

Aishwara Maroju et al. [16] observed that the proposed solution initially yielded poor accuracy, but captions generated after 20 epochs showed relevance. Further training up to 50 epochs significantly improved accuracy and caption quality.

Yan Chu et al. [17] proposed an automatic image captioning methodology using ResNet50 and LSTM, with the AICRL model comprising one encoding and one decoding device. ResNet50 generates a fixed-length vector, which is then used by the LSTM decoder to predict sentences, focusing on specific image areas. The MS-COCO dataset was employed to train the AICRL model.

Shruti Mundargi et al. [18] found that the proposed solution successfully identified crucial elements in images, leveraging ensemble learning to combine outputs from multiple machine learning models, enhancing overall performance.

Janvi Jambhale et al. [24] proposed a Res-Net-LSTM neural network-based model for object recognition and image description. The model uses Res-Net for encoding and LSTMs for decoding. After extracting image features through Res-Net, the model is trained with the extracted features and caption data to construct a vocabulary. Although early training epochs yielded unrelated captions with low accuracy, captions generated after 20 epochs showed some relevance to the test images, with improved accuracy and relevance observed after 50 epochs.

Abhishek Sethi et al. [25] acknowledged that their model's results were not accurate due to language nuances, as they utilized the flickr8k dataset and its Hindi version, which contains English captions translated into Hindi. Despite the challenges in translation due to grammatical and vocabulary differences between English and Hindi, the model was trained on the provided captions and produced results accordingly.

Conclusion

Automated image analysis has the potential to revolutionize healthcare in resource-constrained countries by detecting precancerous and cancerous lesions earlier. The CNN-LSTM architecture is a powerful tool for computer vision and natural language processing, enabling advanced image analysis and captioning capabilities. By leveraging ResNet-LSTM, images can be automatically analyzed and captioned in languages like English, with higher accuracy than comparable models. This technology has vast potential for future image captioning applications, enabling the efficient generation of concise and informative descriptions for large image datasets. AI-powered image captioning can extract detailed textual information from images, enhancing our ability to analyze and understand visual data.

REFERENCES

[1] Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. 2014 .Multimodal neural language models. In Proceedings of the 31st International Conference on Machine Learning (ICML-14).595–603.

[2] Chen X, Fang H, Lin TY, Vedantam R, Gupta S, Dollár P, Zitnick CL (2015) Microsoft coco captions: data collection and evaluation server. arXiv preprint arXiv:1504.00325

[3]. Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. 2015. Deep captioning with multimodal recurrent neural networks (m-rnn).In International Conference on Learning Representations (ICLR).

[4] Soh, Moses. "Learning CNN-LSTM architectures for image caption generation." Dept. Comput. Sci., Stanford Univ., Stanford, CA, USA, Tech. Rep 1 (2016)

[5] Wang, Cheng, Haojin Yang, Christian Bartz, and Christoph Meinel. "Image captioning with deep bidirectional LSTMs." In Proceedings of the 24th ACM international conference on Multimedia, pp. 988-997. 2016.

[6] Justin Johnson, Andrej Karpathy ,and LiFei Fei.2016. Densecap: Fully convolution allocalization networks
fordensecaptioning.InProceedingsoftheIEEEConferenceonComputerVisionandPatternRecognition.4565–4574.

[7] Sun, Qing, Stefan Lee, and Dhruv Batra. "Bidirectional beam search: Forward-backward inference in neural sequence models for fill-in-the-blank image captioning." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6961-6969. 2017

[8] Bin, Yi, Yang Yang, Fumin Shen, Ning Xie, Heng Tao Shen, and Xuelong Li. "Describing video with attention-based bidirectional LSTM." IEEE transactions on cybernetics 49, no. 7 (2018): 2631-2641.

[9] Hossain, Md Zakir, Ferdous Sohel, Mohd Fairuz Shiratuddin, Hamid Laga, and Mohammed Bennamoun. "Bi-SAN-CAP: Bidirectional self-attention for image captioning." In 2019 Digital Image Computing: Techniques and Applications (DICTA), pp. 1-7. IEEE, 2019.

[10] Zhu, Xinxin, Lixiang Li, Jing Liu, Ziyi Li, Haipeng Peng, and Xinxin Niu. "Image captioning with triple-attention and stack parallel LSTM." Neurocomputing 319 (2018): 55-65.

[11] Cui W, Wang F, He X, Zhang D, Xu X, Yao M et al (2019) Multi-scale semantic segmentation and spatial relationship recognition of remote sensing images based on an attention model. Remote Sens 11(9):1044

[12] Subash, R., R. Jebakumar, Yash Kamdar, and Nishit Bhatt. "Automatic image captioning using convolution neural networks and LSTM." In Journal of Physics: Conference Series, vol. 1362, no. 1, p. 012096. IOP Publishing, 2019.

[13] Ignatious, L. Abisha Anto, S. Jeevitha, M. Madhurambigai, and M. Hemalatha. "A Semantic Driven CNN–LSTM Architecture for Personalised Image Caption Generation." In 2019 11th International Conference on Advanced Computing (ICoAC), pp. 356-362. IEEE, 2019.

[14] Sharma, Himanshu, and Anand Singh Jalal. "Incorporating external knowledge for image captioning using CNN and LSTM." Modern Physics Letters B 34, no. 28 (2020): 2050315.

[15] Geetha, G., T. Kirthigadevi, G. Godwin Ponsam, T. Karthik, and M. Safa. "Image captioning using deep convolutional neural networks (CNN)." In Journal of Physics: Conference Series, vol. 1712, no. 1, p. 012015. IOP Publishing, 2020.

[16] Alabduljabbar, Ghadah, Hafida Benhidour, and Said Kerrache. "Image Captioning based on Feature Refinement and Reflective Decoding." arXiv preprint arXiv:2206.07986(2022).

[17] Peng, Yuqing, Xuan Liu, Weihua Wang, Xiaosong Zhao, and Ming Wei. "Image caption model of double LSTM with scene factors." Image and Vision Computing 86 (2019): 38-44.

[18] Liu, M., Li, L., Hu, H., Guan, W., & Tian, J. (2020). Image caption generation with a dual attention mechanism. Information Processing & Management, 57(2), 102178.

[19] Shobiya, L., and R. Pradheesha. "ImageCaption Generator Using RESNET-LSTM" International Journal of Research in Engineering and Science (IJRES) ISSN (Online): 2320-9364.

[19] Chu, Yan, Xiao Yue, Lei Yu, Mikhailov Sergei, and Zhengkui Wang. "Automatic image captioning based on ResNet50 and LSTM with soft attention." Wireless Communications and Mobile Computing 2020 (2020): 1-7.

[20] Mundargi, Shruti, and Hrushikesh Mohanty. "Image Captioning using Attention Mechanism with ResNet VGG and Inception Models." International Research Journal of Engineering and Technology (IRJET) 7, no. 09 (2020).

[21] Aishwarya Maroju, Sneha Sri Doma, Lahari Chandarlapati, 2021, "Image Caption Generating Deep Learning Model", International Journal of Engineering Research & Technology (IJERT) Volume 10, Issue 09 (September 2021).

[22] Jin, Xin, JianwenLv, Xinghui Zhou, Chaoen Xiao, Xiaodong Li, and Shu Zhao. "Aesthetic image captioning on the FAECaptions dataset." Computers and Electrical Engineering 101 (2022): 107866. [14] Li,

[23] Xianrui, Zhiling Ye, Zhao Zhang, and Mingbo Zhao. "Clothes image caption generation with attribute detection and visual attention model." Pattern Recognition Letters 141 (2021): 68-74.

[24] Jambhale, Janvi, Shreeya Sangale, Aarti Avhad, Payal Vairagade, and Jameer Kotwal. "Image caption generator using convolutional neural network and long short-term memory." International Research Journal of Modernization in Engineering Technology and Science(2022).

[25] Sethi Sethi, Abhishek, Aditya Jain, and Chhavi Dhiman. "Image Caption Generator in Hindi Using Attention." In Advanced Production and Industrial Engineering, pp. 101-107. IOS Press, 2022

[26] Antonio M. Rinaldi et al, “Automatic image captioning combining natural language processing and deep neural networks”, Science Direct, Volume 18, June 2023, https://doi.org/10.1016/j.rineng.2023.101107

[27] Gerard Deepak et. al., “Automatic image captioning system using a deep learning approach”, Soft Computing, Published: 27 May 2023