PAPER NAME

Raja Paper Final version (1).docx

AUTHOR

GOPALAKRISHNAN S

WORD COUNT

3297 Words

CHARACTER COUNT

19876 Characters

PAGE COUNT

8 Pages

FILE SIZE

2.2MB

SUBMISSION DATE

Feb 10, 2024 2:58 PM GMT+5:30

REPORT DATE

Feb 10, 2024 2:58 PM GMT+5:30

1% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 0% Internet database
- 1% Publications database
- Crossref database
- Crossref Posted Content database
- 0% Submitted Works database

Excluded from Similarity Report

- Bibliographic material
- Small Matches (Less then 10 words)

Unveiling the Narrative Frontier: Deep Learning-Powered Caption Generation in Machine Vision

Gopalakrishnan S[1] Ramyasree C[1] Kalpana K[1] Lakshitha K[1]

[1]Department of Computer Science and Engineering (Data Science)
gopal.pgsk@gmail.com
ramyacheekala9@gmail.com
kurimisettykalpana@gmail.com
Kollalakshitha123@gmail.com

Madanapalle Institute of Technology & Science, Madanapalle, Andhra Pradesh

ABSTRACT

Developing image captioning models capable of understanding complex contextual details and relationships between objects poses a significant challenge. Current approaches frequently encounter difficulty in discerning intricate contextual cues, resulting in less accurate captions. Overcoming this challenge involves exploring how advanced neural networks can analyze visual content, extract key elements and relationships, and transform them into compelling textual descriptions. This paper addresses these hurdles by leveraging advanced neural networks in deep learning, specifically VGG16 for encoding and Bi-LSTM for decoding. VGG16 captures rich hierarchical features, enhancing the model's understanding of complex visual scenes. The Bi-LSTM architecture aids in capturing both past and future dependencies in sequences, facilitating the generation of coherent and contextually relevant captions. The synergy between VGG16 and Bi-LSTM enables effective navigation of spatial hierarchies in visual data and temporal dependencies in language. The model's performance is impressive, with a BLEU score of 0.51 for the Flickr 8K dataset, emphasizing its ability to capture detailed context effectively. This paper extends to diverse applications, including multimedia content creation.

Keywords : Deep Learning, VGG16, Bi-LSTM, Image Captioning, BLEU score ,Multimedia Content Creation, Visual Content Analysis, Hierarchical Features.

1. INTRODUCTION

In recent times, the surge in digital data on the World Wide Web is exemplified by Flickr, hosting over 3 billion photographs and adding 2.5 million new images daily. Online news sources like CNN, Yahoo!, and the BBC incorporate photos, emphasizing the significance of visual elements in digital communication [1]. Advancements in artificial intelligence (AI) extend to diverse domains, including creating artwork such as pictures, narratives, and movie settings [2].

Picture captioning is a critical focus, addressing challenges in learning the relationship between text and image modalities using computer vision algorithms and natural language processing [3]. Inspired by young children's innate ability to communicate before literacy, there's a growing emphasis on creating environments for AI to understand and describe images effectively [4]. Developments in remote sensing have mostly focused on scene classification, object detection, and segmentation. However, there is a growing possibility to use image captioning methods, which make use of advances in natural language processingand computer vision to provide informative captions[5].

Interpreting visual descriptions becomes paramount with the daily flood of images, necessitating automatic captions for efficient indexing and searches [6].Machines are able to recognize the human activities in videos to a certain extent , but the automatic description of visual scenes has remained unsolved[7].

In order to get over these drawbacks, our suggested model combines the benefits of VGG16 and Bi-LSTM with the goal of improving comprehension of intricate visual sceneries and making it easier to create captions that make sense within their context. For feature extraction from images, the deep convolutional neural network architecture VGG16 functions as a pre-trained model. Rich hierarchical features can be captured with ease thanks to its well-established usefulness in image classification applications. These characteristics help the model better understand complex visual scenes by enabling it to identify important components and connections between pictures.

The Bi-LSTM architecture is essential for capturing past and future dependencies in sequences,

supporting VGG16. The model's ability to comprehend context is improved by this bidirectional processing of sequences, which makes it possible for it to provide more varied and contextually rich captions. Thus, the synergy between VGG16 and Bi-LSTM facilitates efficient navigation of temporal dependencies in linguistic and spatial hierarchies in visual data, opening the door to a more complex and all-encompassing picture captioning system.

The BLEU score, a commonly used metric for evaluating the calibre of machine-generated text, is used to quantify the performance of the suggested model. Our model's impressive BLEU score of 0.51 on the flickr 8k dataset highlights its effectiveness in capturing detailed context. Our method's versatility goes beyond captioning to a wide range of applications, such as multimedia content production and maybe altering the way visually impaired people engage with visual content.

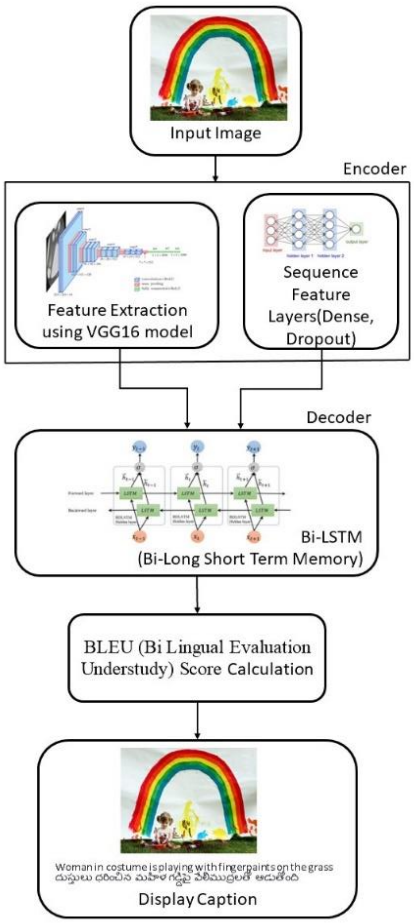
2. LITERATURE SURVEY

In the realm of image captioning, various researchers have proposed innovative frameworks and techniques to enhance the generation of descriptive captions for images.Soheyla Amirian et al. [8] propose a novel framework using scene graph generation for generating summarised captions for image collections. Their method combines scene graphs into a summarised structure, showcasing superior performance over baseline models. Huawei Zhang et al. [9] present a Bi-LSTM model with a subsidiary attention mechanism, addressing inconsistent semantic information and optimizing CIDEr scores through reinforcement learning.

Reshmi Sasibhooshan et al. [10] explore diverse methods for image captioning, including wavelet decomposition and contextual spatial relation extraction. Their proposed models exhibit improved performance and meaningful caption generation. Dinesh Naik et al. [11] contribute to the literature with a framework for video captioning, outperforming existing methods and utilizing visual encoders, stacked LSTM, attention mechanisms, and GloVe word embeddings.

Yana Zhang et al. [12] introduce a Visual Semantic Attention Model (VSAM) for generating visual keywords, showcasing precision and recall improvements in image captioning. Matteo Stefanini et al. [3] present the CPRC approach for semi-supervised image captioning, achieving significant improvements. Additionally, Oriol Vinyals et al. [13] presented a novel method for creating image captions by combining machine translation and computer vision techniques. This technique combines the use of RNN for modeling languages with CNN for picture interpretation. While the RNN component creates meaningful captions according to these qualities, the CNN component evaluates the visual input in order to extract pertinent information from the images.

Rita Ramos et al. [14] propose a novel approach for remote sensing image captioning, surpassing traditional methods and providing open-source code for future research. Another work by Matteo Stefanini et al. [15] covers diverse aspects of deep learning-based image captioning, including evaluation metrics, sub-tasks, training strategies, datasets, and future research directions. Guiguang Ding et al. [16] address image captioning challenges through the R-LSTM method, achieving notable improvements in CIDEr scores on MS COCO and Flickr30k.



3. METHODOLOGY

3.1 Architecture

Figure 1 : Proposed model architecture

3.2 Data Handling and Download:

In this paper, we present the Kaggle API, an effective tool for getting and retrieving datasets

from Kaggle. We make the Flickr 8k dataset easier to retrieve by using the!kaggle datasets download command. This is a popular dataset that is useful for developing image captioning models because it aligns images with written descriptions. By running this command, we improve the dataset retrieval process, which in turn makes the photos and their captions accessible for viewing after the images are extracted from the zip file.

3.3 Knowing about Dataset:

We provide an extensive framework collection that utilizes an 8,000 image selection that has been carefully selected and retrieved from Kaggle in order to facilitate caption-based image search and description. This dataset, referred to as "Flickr 8k", is painstakingly assembled from manually selected images taken from six different Flickr groups, carefully chosen to capture a wide range of events and situations.The dataset has been purposefully filtered to remove any identifying well-known people or objects.Out of all the datasets that are accessible, two standout resources stand out as being particularly useful for picture caption training in computer vision research: Flickr 8k and Flickr 30k. With 8,000 and 31,000 photos, respectively, each image in both datasets is accompanied by five different explanations, adding significant contextual knowledge to the collection.Given the limitations imposed by finite storage and processing power, our system makes use of the Flickr 8k dataset first, taking advantage of its comparatively lower number of images for more efficient testing. We use Andrej Karpathy's painstakingly created, freely available, and cleaned Flickr 8k data break as our main input dataset in order to guarantee consistency and improve usability. Among the crucial changes this preprocessing stage includes are lowering the primary Flickr 8k text to smaller letters, removing non-alphanumeric letters, and dividing the dataset into separate train and test subsets.

3.4 Data Preprocessing:

The VGG16 model is utilized in this paper to extract features from images, and its preprocessing function is used to adaptively alter the images as needed. Then, we use the pickle module to categorize and preserve the attributes that we have gathered, making them available for use in the following stages of our process.

3.4.1 Image Preprocessing:

Adhering to a standardized approach when supplying images into the VGG16 model post-preprocessing is prioritized, ensuring uniformity and consistency in their format.In order to achieve this goal, we use methods within the Keras library for loading and scale images to a standard size that is predetermined, usually 224 by 224 pixels. Furthermore, the required process of normalizing the values of pixels is carried out in order to comply with the requirements of the model VGG16, which normally works on images with standardized pixel values. Furthermore, we guarantee that any image enhancement methods used in preprocessing preserve data integrity while improving the generalization of models abilities. This methodical approach to preprocessing guarantees comply with the VGG16 model's specifications, which improves accuracy and performance in later processing stages.

3.4.2 Caption Preprocessing:

It is highlighted how important it is to preprocessing on captions to improve the model's capacity to learn from written input. Lowercasing ensures the text is consistent, and removing unusual characters, noise, and digits speeds up data processing. We also provide beginning and ending tags during training, which are important for helping the model understand the borders of a caption.

3.5 Image Feature Extraction:

The application of the VGG16 algorithm to feature extraction from pictures is investigated. We preprocess the photos to make the necessary adjustments for feature extraction, making use of the capabilities of the model. We next use the pickle module to classify and store the extracted characteristics so that they can be used for additional processing and analysis in later phases of the study.

3.5.1 VGG16 Model:

Feature extraction is achieved by a modification of the popular model VGG16, which is widely recognized for its effectiveness in picture classification tasks. We remove the last classification layer from the model and refocus it to extract rich, primary characteristics from images. By capturing crucial visual information, our enhanced VGG16 model produces an output that essentially serves as a brief summary of the input pictures. By removing the last classification layer, the model is ensured to focus on retrieving rich, high-level properties from images. This revised VGG16 model produces an output that serves as a streamlined version of the original input images and stores significant visual information.

3.6 Tokenizer:

Our model uses the Keras library's Tokenizer class to turn processed caption into numerical data. The process of tokenization, which involves giving a distinct integer to every distinct term in the lexicon, is essential for transforming text into a format that can be fed to our model for processing. In order to guarantee that textual data will be usefully employed in our study going forward, this step is crucial.

3.7 Model Architecture:

we detail the design and structure of our neural network model. Our model architecture encompasses the arrangement and configuration of various layer that are mentioned below.

3.7.1 Encoder:

In this study, we extract visual features by using the upgraded VGG16 model. The encoder component of our model then processes these features. These visual elements gain representational strength as they move through a thick layer.

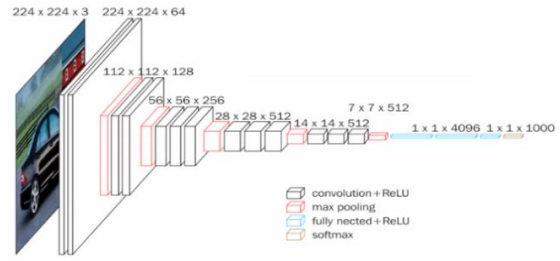


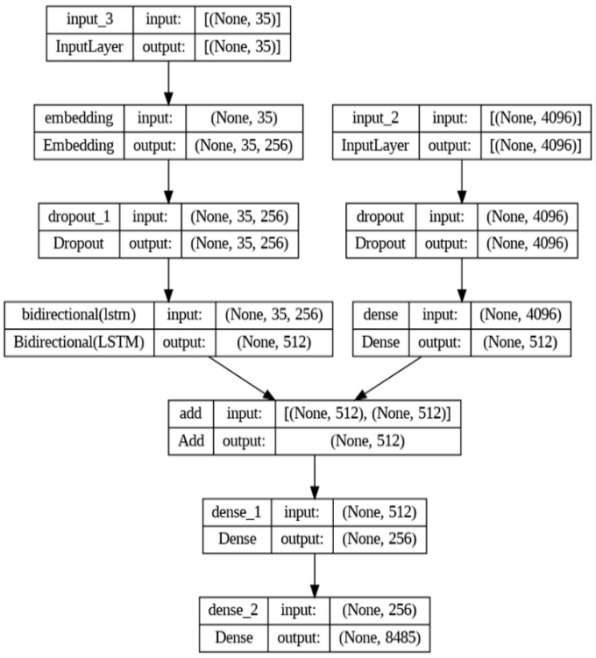
Figure 2: Overview of VGG16 model layers.

3.7.2 Decoder:

The decoder part of our model combines the sequence characteristics and augmented images. After then, further deep layers of processing are applied to these merged features. The final layer of our model predicts word probabilities using a softmax activation. This output shows our model's recommendation for the next word in the sequence, which completes the image caption.

3.7.3 Compilation:

The Adam optimizer and categorical cross-entropy loss are used in the compilation of our model. For multi-class classification problems, we find that logistic cross-entropy is an excellent choice. Additionally, we leverage the Adam optimizer due to its flexible learning rate capabilities, which greatly facilitate training convergence.



We also build sequence feature layers. Sequence feature layers include LSTM, Dropout, and Dense layers, among others. The encoder is followed by the decoder, which creates the model using dense layers. We train our model by specifying parameters such as epochs and batch size. We can easily train the model as we already have a data generator that requires minimal memory. After training the model, we can save it. We generated a new model by changing additional layers. Batch normalisation and layer normalisation are similar. Batch normalisation is widely utilised, especially in CNNs. Although the input images have previously been processed, training changes the parameters, requiring a new normalisation of the layer inputs.

Figure 3: Implementation of the model

3.8 Training:

3.8.1 Custom Data Generator:

To overcome any possible memory constraints, we developed a special data generator. We use this generator to efficiently handle batch input of pictures and captions during training. This method facilitates seamless model training and avoids memory fatigue.

3.8.2 Epochs and Training:

To train the model for this study, we use 50 epochs, which correspond to whole dataset iterations. To minimize the provided categorical cross-entropy loss, we use the fit technique to modify the weights of the model during training. By going through this

iterative process, we improve the model's caption accuracy.

3.9 Evaluation:

3.9.1 BLEU Score:

We apply a widely-used measure in machine translation jobs, the BLEU score, to comprehensively assess the quality of the automatically generated captions. It evaluates how close to the real titles in the test data and the ones produced by our model are to each other. This review helps us understand how well our algorithm works to provide meaningful and accurate captions.More specifically, the produced and reference captions showed a reasonable degree of alignment, as indicated by our BLEU score of 0.51.

3.10 Caption Generation and Translation:

3.10.1 Generating Captions:

With the model that has been trained, we generate test set images with specific descriptions. The generated captions, which make use of the learnt features, provide information about how well the model was able to communicate visual content.

3.10.2 Translation:

With the translate library, we further translate generated captions into Telugu. This step extends the model's utility to a multilingual setting and allows us to examine its cross-lingual capabilities.

3.10.3 Text-to-Speech:

To turn the translated text into speech, we use the Google Text-to-Speech (gTTS) API. By taking this extra step, we offer an aural depiction of the model's output, improving the user experience's accessibility and immersion.

3.11 Visualization:

The flexible charting program Matplotlib is used to show the source images. We use this data visualization tool to assess our model's performance on a qualitative level. Furthermore, we print both the generated and the actual captions, enabling a thorough examination of the generated textual outputs. By means of these illustrations and written explanations, we enable an extensive analysis of the model's functionality and produced results.

4. RESULT ANALYSIS

Images with their generated captions, both in English and Telugu as predicted by the system, are presented in

Figure 4 - 7

Figure 4:

White dog is running through the snow

తెల్లకుక్కమంచుగుండాపరుగెత్తుతోంది

Boy in red shirt kneeling on thin stick

ఎరుపుచొక్కాలోఉన్నబాలుడుసన్నని కర్రపైమోకరిల్లుతున్నాడు

Figure 6:

Figure 5:
man and woman are paddling boat through water
పురుషుడుమరియుస్త్రీనీటిద్వారాపడవనుతెడ్డువే
స్తున్నారు

Figure 7:
young boy is hanging upside down on swing
యువకుడుస్వింగ్
లోతలక్రిందులుగావేలాడుతున్నాడు

Table 1: Results comparison

Model	BLEU(Bi Lingual Evaluation Understudy)
CNN + Bi - LSTM	0.37
Inception V3 + LSTM	0.50
VGG16 + Bi - LSTM (Our Model)	0.51

The provided tables indicate that the VGG16+Bi-LSTM model achieves the highest score, followed by the Inception V3 + LSTM and CNN+Bi-LSTM models. Our analysis reveals that our VGG16+Bi-LSTM model outperforms the others. In summary, these tables emphasize the superiority of our VGG16+Bi-LSTM model in producing accurate and fluent translations.

5. CONCLUSION

In conclusion, by combining VGG-16 for image encoding and Bidirectional LSTM for caption decoding, the Image Caption Generatorpioneers deep learning. It meets obstacles head-on and establishes new benchmarks for comprehending intricate visual content. Throughout the study, we demonstrated how well deep learning with VGG-16 extracts important visual characteristics, and how accurate captions are produced by the bidirectional LSTM, which captures spatial hierarchies and temporal connections. Our method improves image captioning by highlighting the complementary work that computer vision and natural language processing can do together. Though significant progress has been made, more study may concentrate on optimising designs, adjusting parameters, and investigating multimodal learning strategies for complex interactions. Beyond merely improving technical aspects, our work gives robots the ability to produce precise captions, providing content creators with strong communication tools. The initiative has the potential to improve accessibility and give those with visual impairments a more meaningful way to interact with visual content.

6. FUTURE WORK

In the realm of image caption generation, future work holds exciting possibilities. Model refinement and optimization suggest fine-tuning existing models for enhanced performance and tailoring them to specific domains like medical or artistic images. Attention mechanisms could elevate contextual understanding by allowing the model to focus on relevant image regions. Multimodal learning opens avenues for fusing text and image information, extending to audio-visual captioning for comprehensive multimedia understanding. Fine-grained captioning delves into object-centric details and complex relationships. Transfer learning explores broader datasets and cross-domain adaptations. Improved evaluation metrics and also human-centric assessments ensure quality captions. Interactive and adaptive systems enable user interaction and feedback adaptation. Ethical considerations address biases and potential misuse.Real-time applications optimise speed, and cross-modal retrieval systems expand utility by retrieving images based on textual queries and vice versa.

7. REFERENCES

[1] YANSONG FENG and MIRELLA LAPATA, "Automatic Caption Generation for News Images", IEEE, APRIL 2013.

[2] KYUNGBOK MIN, MINH DANG and HYEONJOON MOON, "Deep Learning-Based Short Story Generation for an Image Using the Encoder-Decoder Structure",IEEE Access, August 19, 2021.

[3] YANG YANG ,HONGCHEN WEI, HENGSHU ZHU,DIANHAI YU,HUI XIONG and JIAN YANG ,"Exploiting Cross-Modal Prediction and Relation Consistency for Semisupervised ImageCaptioning", IEEE, FEBRUARY 2024.

[4] JOHANES EFFENDI , SAKRIANI SAKTI and SATOSHI NAKAMURA , (Fellow, IEEE),"End-to-End Image-to-Speech Generation for Untranscribed Unknown Languages", IEEE Access, April 15, 2021.

[5] XIAOQIANG LU , SENIOR MEMBER,BINQIANG WANG, XIANGTAO ZHENG and XUELONG LI , "Exploring Models and Data for Remote Sensing Image Caption Generation", IEEE , APRIL 2018.

[6] MR.N. RAGHU,SAI SRIKAR, AAFTAAB and RUTHVIK SAI," IMAGE CAPTIONING USING DEEP LEARNING ", IJRTI, 2023.

[7] SHENG LI , ZHIQIANG TAO , KANG LI, and YUN FU , "Visual to Text: Survey of Image and Video Captioning", IEEE, AUGUST 2019.

[8] SOHEYLA AMIRIAN, KHALED RASHEED, THIAB R. TAHA and HAMID R. ARABNIA ,"Automatic Image and Video Caption Generation With Deep Learning: A Concise Review and Algorithmic Overlap", IEEE Access, December 16, 2020.

[9] KAPIL HANDE, HRUSHI KARLEKAR, PRANIT YEOLE, ADITYA LIKHAR and HIMANSHU RANGARI, "NLP based Video Summarisation using Machine Learning", International Journal of Scientific Research in Science, 14 April 2023.

[10] DINESH NAIK and C. D. JAIDHAR " A novel Multi-Layer Attention Framework for visual description prediction using bidirectional LSTM ", Spinger Open, 2022.

[11] RESHMI SASIBHOOSHAN,SURESH KUMARASWAMY and SANTHOSHKUMAR SASIDHARAN ," Image caption generation using Visual Attention Prediction and Contextual Spatial Relation Extraction", Spinger Open, 2023.

[12] YANA ZHANG, ZEYU CHEN and ZHAOHUI LI " VSAM-Based Visual Keyword Generation for Image CaptionSUYA ZHANG", IEEE Access, February 19, 2021.

[13] ORIOL VINYALS,ALEXANDER TOSHEV,SAMY BENGIO and DUMITRU ERHAN,"Show and Tell: A Neural Image Caption Generator", Computer Vision Foundation, 2015.

[14] RITA RAMOS and BRUNO MARTINS ,"Using Neural Encoder-Decoder ModelsWith Continuous Outputs for Remote Sensing Image Captioning" ,IEEE Access, March 9, 2022.

[15] MATTEO STEFANINI, MARCELLA CORNIA , LORENZO BARALDI , SILVIA CASCIANELLI ,GIUSEPPE FIAMENI and RITA CUCCHIARA ,"From Show to Tell: A Survey on Deep Learning-Based ImageCaptioning", IEEE, JANUARY 2023.

[16] GUIGUANG DING,MINGHAI CHEN,SICHENG ZHAO,HUI CHEN,JUNGONG HAN and QIANG LIU , “Neural Image Caption Generation with Weighted Training and Reference", Springer, 2019.

1% Overall Similarity

Top sources found in the following databases:

- 0% Internet database
- Crossref database
- 0% Submitted Works database
- 1% Publications database
- Crossref Posted Content database

TOP SOURCES

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	**Sheng Li, Zhiqiang Tao, Kang Li, Yun Fu. "Visual to Text: Survey of Ima...** Crossref	<1%