iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VI **Month of publication:** June 2023

DOI: https://doi.org/10.22214/ijraset.2023.53825

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VI Jun 2023- Available at www.ijraset.com

Image Caption Generator Using Attention Based Neural Networks

Ashwini Mahendiran[1], Dr. Muthuram. R[2]
[1]*P.G. Student, Department of Computer Science, Government College Of Technology, Coimbatore, Tamil Nadu, India*
[2]*Associate Professor, Department of Computer Science, Government College Of Technology, Coimbatore, Tamil Nadu, India*

Abstract: ***Image caption generation is a method used to create sentences that describe the scene depicted in a given image. The process includes identifying objects within the image, carrying out various operations, and identifying the most important features of the image. Once the system has identified this information, it generates the most relevant and concise description of the image, which is both grammatically and semantically correct. With the progress in deep-learning techniques, algorithms are able to generate text in the form of natural sentences that can effectively describe an image. However, replicating the natural human ability to comprehend image content and produce descriptive text is a difficult task for machines. The uses of image captioning are vast and of great significance, as it involves creating succinct captions utilizing a variety of techniques such as Natural Language Processing (NLP), Computer Vision(CV), and Deep Learning (DL) techniques. The current study presents a system that employs an attention mechanism, in addition to an encoder and a decoder, to generate captions. It utilizes a pre-trained CNN, Inception V3, to extract features from the image and a RNN, GRU, to produce a relevant caption. The attention mechanism used in this model is Bahdanau attention, and the Flickr-8Kdataset is utilized for training the model. The results demonstrate the model's capability to understand images and generate text in a reasonable manner.***
Keywords: Machine Learning, Deep Learning, Image captioning, attention, image processing.

I. INTRODUCTION

The significance of robots interpreting images through image captioning is highlighted in this text. It entails locating objects in an image and determining their characteristics and interactions. Assistive technologies, human-computer interfaces, and image search engines can all benefit from image captioning. The architecture consists of a language decoder for caption generation and an image encoder for feature extraction. CNNs and other deep learning models are good at recognizing images. For language modelling in caption generation, RNNs and LSTMs are utilised. Focusing on pertinent visual regions is assisted by attention mechanisms. Accurate object recognition and object information loss during feature learning are problems. It's crucial to use synthetic visuals for training and testing. The goals are to produce captions of the highest calibre, improve deep networks with attention, provide contextual information, and show how helpful synthetic images are.

II. SIGNIFICANCE OF THE SYSTEM

The paper on image caption generation holds significance as it addresses the challenging task of generating descriptive sentences that accurately describe the content of an image. This process involves identifying objects, performing operations, and extracting important features from the image. While deep learning techniques have made progress in generating natural language sentences, replicating human-level comprehension and descriptive abilities remains difficult for machines. Image captioning has wide-ranging applications and is crucial in fields such as Natural Language Processing (NLP), Computer Vision (CV), and Deep Learning (DL). The presented system in the study utilizes an attention mechanism, encoder, and decoder, leveraging a pre-trained CNN (Inception V3) for feature extraction and an RNN (GRU) for caption generation. The use of Bahdanau attention and training on the Flickr-8K dataset demonstrates the model's competence in understanding images and generating coherent textual descriptions.

III. LITERATURE SURVEY

Cho et al. [1] used convolutional neural networks (CNNs) and gated recurrent neural networks (GRNNs) to train encoder-decoder networks with attention mechanisms in their study. For machine translation, they used an attention-based recurrent neural network learning model (RNN-LM) as the decoder and a bidirectional recurrent neural network (BiRNN) as the encoder. Similar to this, Xu et al. proposed two categories of attention-based image caption generators: hard stochastic attention and soft deterministic attention.

©IJRASET: All Rights are Reserved

Their analysis concentrated on how well attention was paid to the "where" and "what" elements of the image when creating image captions. They utilised a CNN to extract picture attributes as input, and an LSTM with the context vector to generate words at each step.

Li et al.[3] study introduced an attention-based scene text recognition model that does not require image segmentation. Their model incorporates sequence recognition through an LSTM network, feature attention through a CNN, and feature extraction through a CNN, all within a cooperatively trainable network. The model was evaluated on the IIIT5K, SVT, ICDA2003, and ICDAR2013 datasets and proved to be more effective than earlier methods, however it still has text recognition issues. Future study should investigate the application of a more complex CNN architecture to improve the model's applicability in word recognition from photos.

Fu et al.[4] developed an automated system for captioning photos that creates precise, pertinent phrases from photographs. By incorporating scene-specific context, the system uses visual perception to construct word sequences that encode the image with higher-level semantic information. Standard datasets including Flickr8K, Flickr30K, and MSCOCO were evaluated using both automatic metrics and human evaluation. Incorporating either scene-specific context or region-based attention increased the system's performance. The combination of these two techniques offers encouraging possibilities for obtaining cutting-edge outcomes in upcoming image captioning assignments.

Bin et al.[6] created an adaptive attention technique for visual captioning that emphasises both verbal proficiency and salient visual material. The method incorporates a visual captioning model, linguistic knowledge embedding, and attribute learning, together with an RNN as an encoder and an LSTM as a decoder, and an adaptive attention mechanism.The method also embeds linguistic knowledge from earlier hidden states through a latent representation and uses a pre-trained multi-label classifier to regulate the visual captioning model through a visual gate. The outcomes of the experiment demonstrate how effective the adaptive attention method is.

To overcome the constraints of long-short-term memory (LSTM) structures, Zhu et al.[5] created the Captioning Transformer (CT) model, which employs stacked attention modules without time dependencies. They also suggested a multi-level supervision training strategy. This model has a Convolutional Neural Network (CNN) encoder that extracts image features using ResNet and ResNext image classification models, and a transformer model with stacking attention mechanism as the decoder that turns image characteristics into sentences. Three techniques are used to include picture features into the transformer model: the image feature in front of text embedding in the spatial image feature map, merging the image feature and each word in the spatial image feature map, and using the image feature in the image spatial feature map.

For captioning purposes, Qu et al.[7] suggested using a visual attention mechanism that is applied to Long Short-Term Memory (LSTM). A CNN is used in the model to extract attributes including colour, size, and location; an LSTM is used to produce phrases; and an attention mechanism is used to describe significant objects in the image. Using VggNet, CNN derives features from the image. The attention mechanism in the LSTM comprises two components: input and forget gates, attend and output gates, and memory cells. colouring stimuli driven and stimulus-driven in dimension. The performance of the proposed model was assessed using industry-accepted assessment criteria like BLEU while it was tested on three well-known benchmark datasets: Flick8k, Flick30k, and MSCOCO. The suggested design can generate more interpretable sentences and achieve higher accuracy in object recognition. Future work should use unsupervised data to gain a more comprehensive and precise understanding of the entire image.

IV. METHODOLOGY

The goal of this study is to develop an image captioning system that utilizes a pre-trained CNN for extracting features from an image, combines those features with an attention mechanism, and generates captions using an RNN. The system utilizes multiple pre-trained CNNs for encoding the image into a feature vector. A GRU-based language model is employed as the decoder to construct the descriptive sentence. Additionally, the Bahdanau attention model is integrated with the GRU to improve performance by focusing on specific parts of the image. The system was evaluated on the MSCOCO dataset and demonstrated competitive results compared to current state-of-the-art methods.

A. System Model

A caption generation model using attention typically includes an encoder and a decoder. The encoder converts an image into a feature representation, and the decoder generates a natural language description of the image based on that feature representation. The attention mechanism is used to selectively focus on different parts of the image during the captioning process.

The image captioning system involves several key steps:

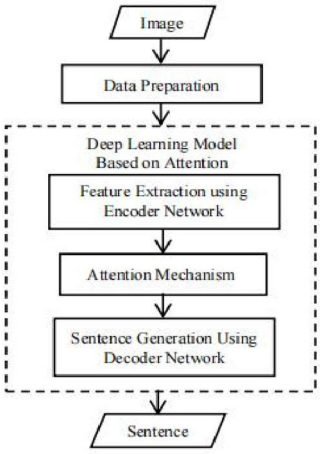


Fig 1. Steps in Image Captioning

The input image is processed through a CNN to extract features and create a feature map.
The encoded features are passed to the RNN-based decoder, such as LSTM, to generate a sequence of words that form the final caption. Additionally, an attention mechanism is utilized to selectively focus on different areas of the image throughout the caption creation process. This mechanism calculates relevance weights for each element in the encoded features, using the current state of the decoder and the encoded features. The weighted sum of the encoded features, known as the context vector, is used as additional input for the LSTM to generate the next word in the caption. This process is repeated until a stopping criterion is met, such as generating an end-of-sentence marker.

1) Data Preparation

Data preparation for attention-based image captioning systems typically involves several steps, such as acquiring and organizing a dataset of images and their corresponding captions, dividing the dataset into training, validation, and test sets, image pre-processing, caption pre-processing, generating captions for the images, and using the attention mechanism to match the captions with the image. It's worth noting that this is a general overview of the data preparation process and specific details may vary depending on the model's architecture and requirements.

2) Feature Extraction Using Encoder Network

The encoder in an attention-based image captioning system is responsible for extracting important features from the input image, a process known as feature extraction. This process typically involves passing the image through a pre-trained convolutional neural network (CNN) which generates a feature map that encodes information about the image at different scales and locations. The feature map is then passed through a pooling layer to reduce dimensionality and capture information at different scales and locations. The final output of this process is a fixed-length feature vector that is used as input to the decoder to generate the final caption. This feature extraction process allows the encoder to create a condensed, yet informative representation of the image which can be used to generate an accurate caption.

3) Attention Mechanism

The attention mechanism in image captioning models is a technique that enables the model to selectively focus on certain parts of an image when generating captions. This helps to create more precise and logical captions by taking into account the most relevant information present in the image. The attention mechanism works by determining the importance of each element in the encoded image features, which are generated by the encoder. These weights are used to create an attention map, which highlights the parts of the image that are most crucial for generating each word in the caption. The attention map is then used to selectively weight the encoded image features, creating a context vector that is used as additional information for the decoder to decide what to generate next.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Different types of attention mechanisms can be used, such as Bahdanau attention mechanism, which uses a neural network to calculate attention weights, or the simpler dot-product attention. These attention mechanisms can be applied in various ways, for example, at each time step of the decoding process or only at specific time steps. In summary, the attention mechanism in image captioning models allows the model to focus on the most important parts of the image during caption generation, resulting in more accurate and coherent captions.

4) Sentence Generation Using Decoder Network

The decoder in an image captioning model is responsible for converting the features extracted from an image into a natural language sentence describing the image. This component of the model typically uses a recurrent neural network such as LSTM or GRU to process the features from the encoder, which is usually a CNN, and generates a sequence of words that make up a sentence. The process of generating the sentence is referred to as "decoding" and it involvespredicting the next word in the sequence based on the previous words and the features from the encoder. Additionally, the decoder can use an attention mechanism which enables it to focus on different regions of the image when generating different parts of the caption.

V. PROPOSED ARCHITECTURE

The process of creating captions for images includes pre-processing the captions and images, using an encoder to extract features from the image, and then using a decoder and an attention mechanism to translate the image into natural sentences.

In this specific approach, the captions were cleaned, a vocabulary was formed, and the sentences were converted into a form that the machine can understand. For the images, they were pre-processed to meet the requirements of the encoder, which in this case is a CNN (InceptionV3). The decoder used is a RNN, specifically a GRU, and it is paired with an attention mechanism to focus on specific regions of the image and produce more accurate descriptions. This is known as an "inject" architecture, where the context vector, after going through the attention mechanism, is injected into the GRU to generate the appropriate words for the caption.

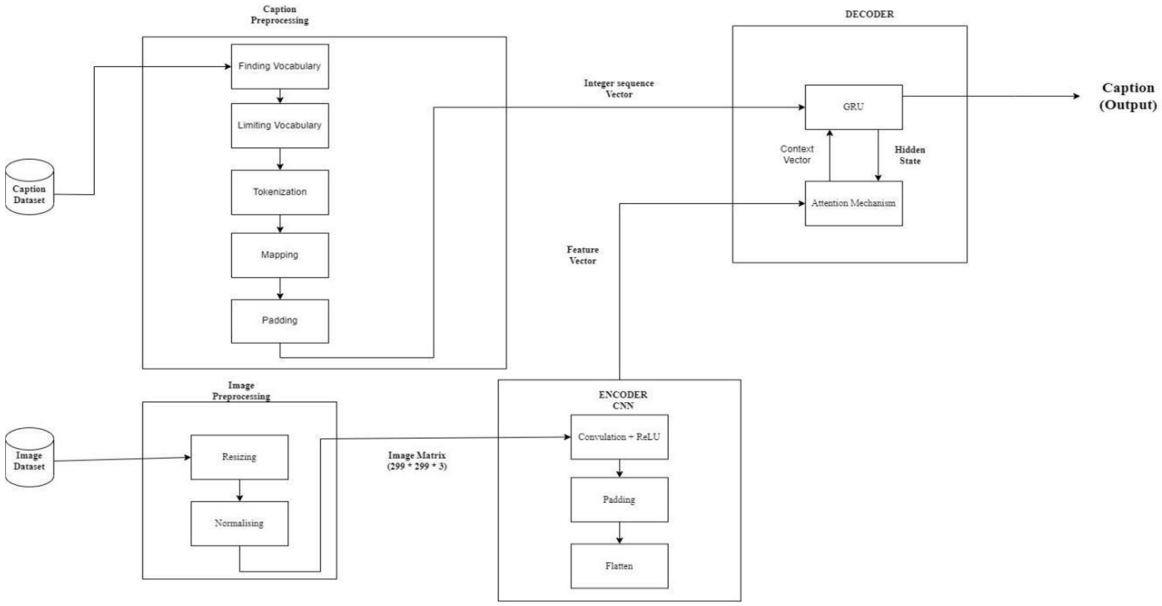


Fig 2. Proposed Architecture

The process of creating captions for images includes pre-processing the captions and images, using an encoder to extract features from the image, and then using a decoder and an attention mechanism to translate the image into natural sentences.

In this specific approach, the captions were cleaned, a vocabulary was formed, and the sentences were converted into a form that the machine can understand. For the images, they were pre-processed to meet the requirements of the encoder, which in this case is a CNN (InceptionV3). The decoder used is a RNN, specifically a GRU, and it is paired with an attention mechanism to focus on specific regions of the image and produce more accurate descriptions. This is known as an "inject" architecture, where the context vector, after going through the attention mechanism, is injected into the GRU to generate the appropriate words for the caption.

A. Deliverables

1) Input – Image

2) Output – Caption for the given image

VI. EXPERIMENTAL RESULTS

A. Quantitative Results

We present results from our proposed technique and compare it to seven state-of-the-art models using multiple metrics. Our technique, NIC, uses image features extracted from a deep CNN and injects them into the first time step of an LSTM-based language model. We compare our results to Soft-Att, MSM, Attribute-driven, NBT, GCN-use LSTM, GCN-use LSTM of visual relationships, and other models. Our results show that the Inception model outperforms the others, followed by ResNet101. However, these results can be improved by using more data for training, as we were limited to 113,287 images due to computational constraints.

Based on the true rate and false rate values of spam and good message, the following graph is generated.

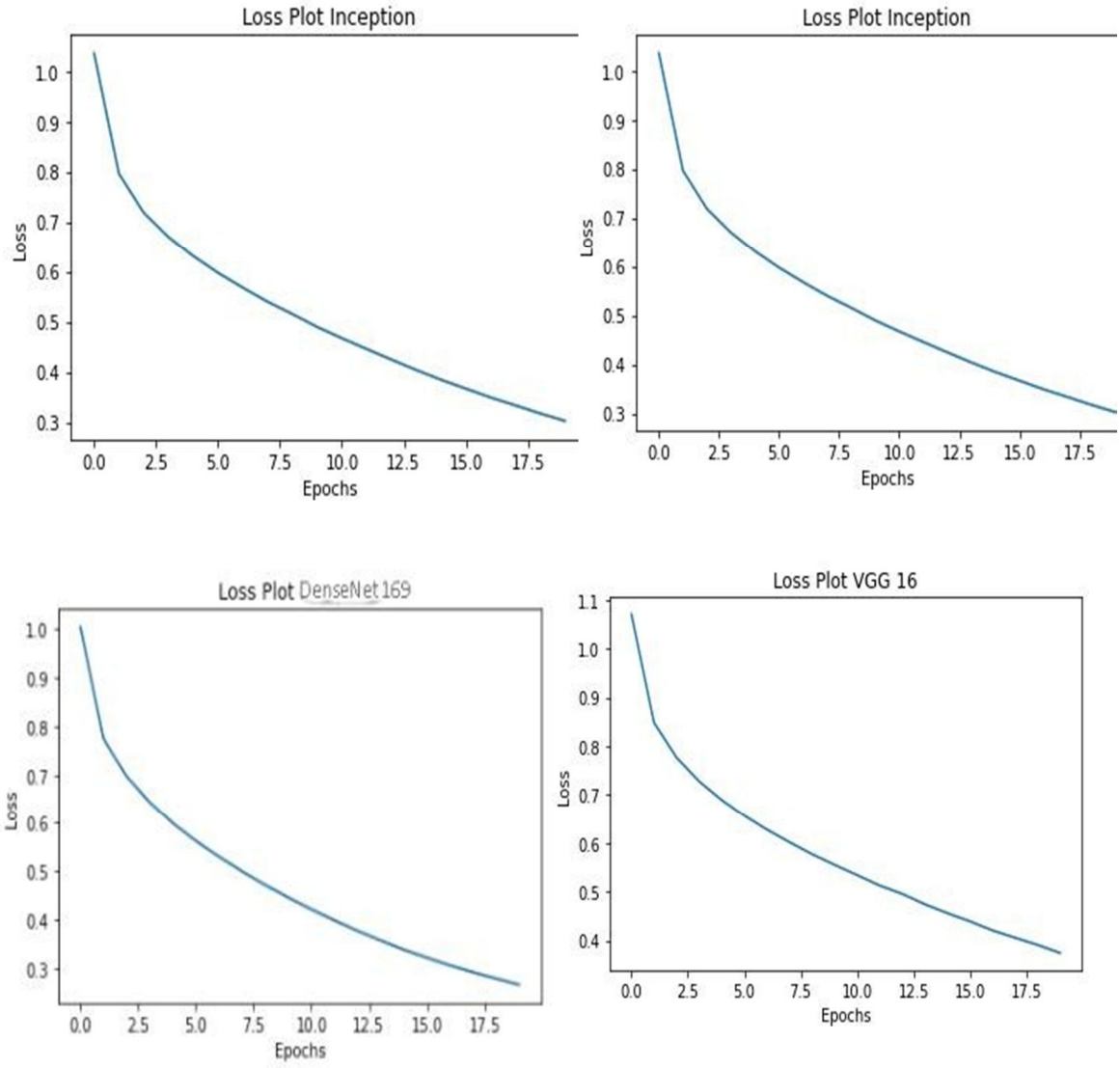


Fig 3. Loss Plot

Table 1: Results

EXPERIMENTAL RESULTS OF STATE-OF-THE_ART MODELS							
MODEL	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	CIDER	METEOR
Google NIC	0.67	0.45	0.30	0.20	--	--	--
Soft Attention	0.71	0.49	0.34	0.24	--	--	0.24
MSM	0.71	0.57	0.43	0.33	0.54	1.02	0.25
Attribute- driven Attention	0.74	0.56	0.44	--	0.55	1.104	--
NBT	0.75	--	0.34	--	--	1.107	0.27
Context-aware Attention	0.76	0.60	0.46	0.36	0.56	1.103	0.28
GCN-LSTM	0.77	--	--	0.3	0.57	1.107	0.28
PERFORMANCE OF OUR PROPOSED GRU ATTENTION-BASED MODELS							
MODEL	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	CIDER	METEOR
Inception V3	0.78	0.57	0.44	0.36	0.59	1.105	0.27
VGG 16	0.74	0.57	0.44	0.33	0.56	1.109	0.2
DenseNet169	0.74	0.56	0.43	0.36	0.58	1.103	0.27
ResNet101	0.75	0.56	0.44	0.37	0.59	1.104	0.29

B. Qualitative Results

The qualitative results of our model are noteworthy. The model generates captions that are coherent and grammatically correct for various images, as shown in Figure 5.2. However, some captions do not accurately describe the image, while others are unreadable. It is difficult to determine whether these errors are a result of poor image recognition or poor text generation.

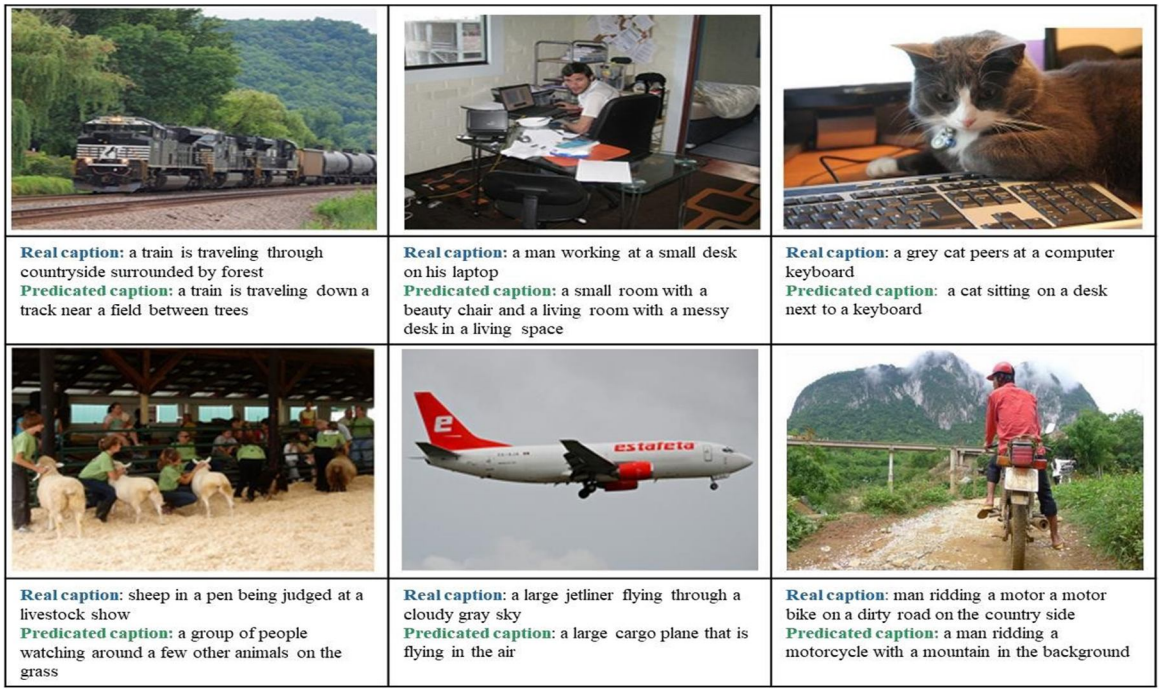


Fig 4. Example captions from conventional mode



Fig 5. model predicts poor captions in this case

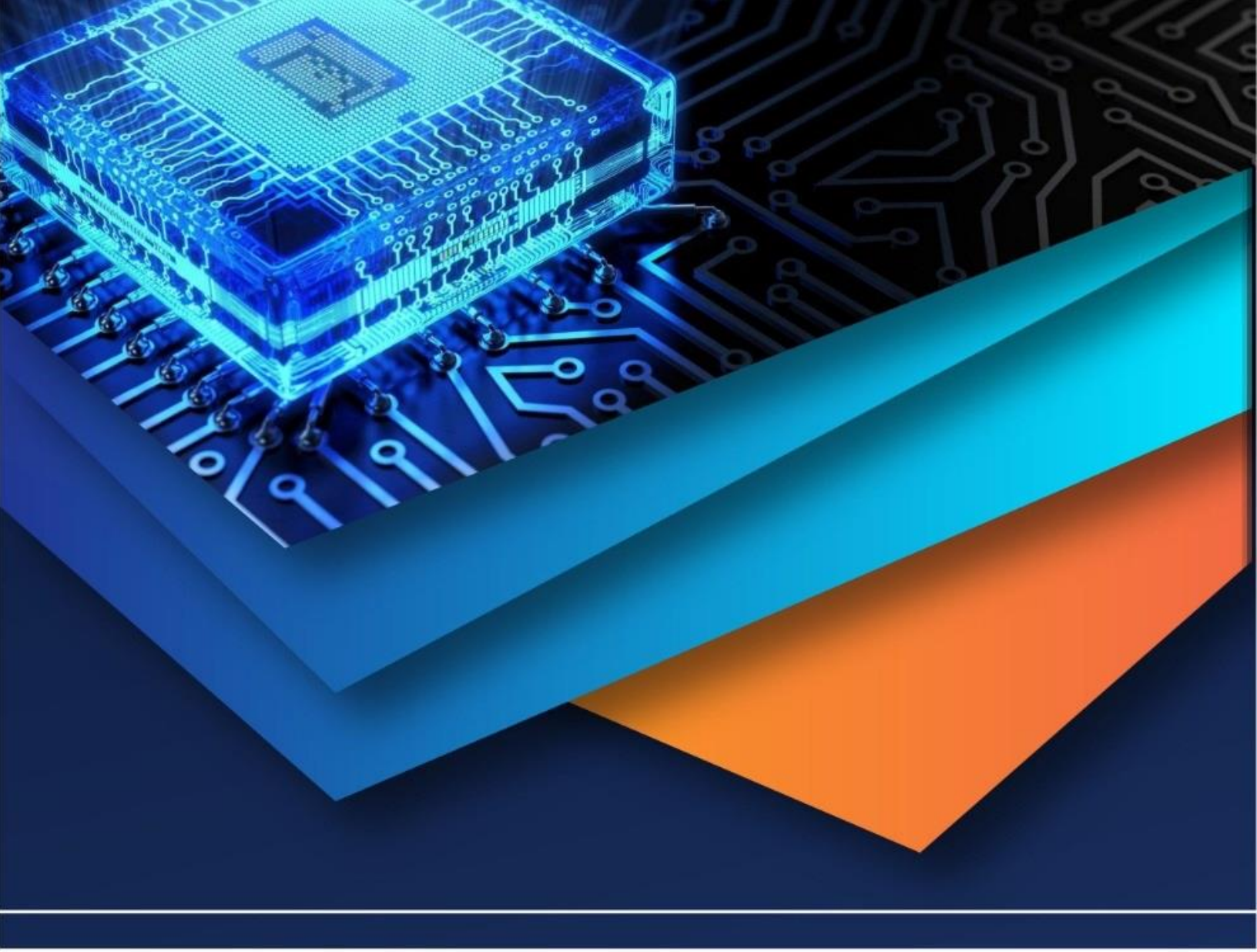
VII. CONCLUSION AND FUTURE WORK

In this study, we proposed an image captioning model that combines the use of a CNN encoder and a GRU decoder with an attention network. The model uses an encoder-decoder architecture, where a pre-trained CNN is utilized to extract features from an image, which is then passed to a GRU-based decoder to generate a descriptive sentence. Additionally, the model incorporates the Bahdanau attention mechanism to focus on specific regions of the image during caption generation, resulting in more accurate and coherent captions. The model can be trained using stochastic gradient descent for ease of use. Results show that the proposed model is effective in generating captions for images.

As a future direction, the proposed image captioning model can be extended by incorporating other computer vision techniques such as object detection and semantic segmentation. By combining these techniques with the current model, it is expected to improve its performance in terms of captioning accuracy and diversity. Object detection techniques, such as YOLO, can be used to detect and identify objects within an image, providing additional information to the model about the image content. Similarly, semantic segmentation techniques, such as FCN, can be used to segment an image into different regions, providing a more detailed understanding of the image content to the model. The combination of these techniques with the current model can help to enhance the ability of the model to capture the fine-grained details of an image, leading to more accurate and comprehensive captions. Further research in this direction can also explore how to effectively combine and fuse information from different computer vision techniques to improve image captioning performance.

REFERENCES

[1] K. Cho, A. Courville, and Y. Bengio, “Describing Multimedia Content Using Attention-Based EncoderDecoder Networks”, IEEE Trans. on Multimedia, 17(11):1875-1886, 2015.
[2] L. Xu, J.L. Ba, R. Kiros, K Cho, A. Courville, R. Salakhudinov, R. S.Zemel and Y. Bengio, “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention”, International conference on machine learning, pp. 2048-2057, 2015.
[3] S. Li, M. Tang, A. Guo, J. Lei, and J. Zhang, “Deep Neural Network with Attention Model for Scene Text Recognition”, IET journals of the institution of Engineering and Technology, 11(7):605-612, 2017
[4] K. Fu, J. Jin, R. Cui, F. Sha, and C. Zhang, “Aligning Where to see and What to Tell: Image Captioning with Region-Based Attention and Scene- Specific Contexts”, IEEE Trans. on Pattern Analysis and Machine Intelligence, 39(12):2321-2334, 2017.
[5] Multimedia Comput. Commun. Appl., 14(3):73, 2018. [21] X. Zhu, L. Li, J. Liu, H. Peng, and X. Niu, “Captioning Transformer with Stacked Attention Model”, Applied Sciences, 8(5):739, 2018
[6] Y. Bin, Y. Yang, J. Zhou, Z. Huang, and H.T. Shen, “Adaptively Attending to Visual Attributes and Linguistic Knowledge for Captioning”, In Proceedings of the 2017 ACM on Multimedia Conference, pp. 1345- 1353, 20
[7] S. Qu, Y. Xi, and S. Ding, “Visual Attention Based on Long-Short Term Memory Model for Image Caption Generation”, Control and Decision Conference (CCDC), 2017 29th Chinese, IEEE, pp. 4789-4794, May 2017.
[8] L. Li, S. Tang, Y. Zhang, L. Deng, and Q. Tian, “GLA: Global-Local Attention for Image Description”, IEEE Trans. on Multimedia, 20(3):726- 737, 2018.
[9] S. Ye, J. Han, and N. Liu, “Attentive Linear Transformation for Image Captioning”, IEEE Trans. on Image Processing, 27(11):5514-5524, 2018.
[10] Cornia, Marcella, et al. “Paying more attention to saliency: Image captioning with saliency and context attention”, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 14.2 (2018): 48.
[11] A. Wang, H. Hu and L. Yang, “Image Captioning with Affective Guiding and Selective Attention”, ACM Trans. Multimedia Comput. Commun. Appl., 14(3):73, 2018.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)