iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** III **Month of publication:** March 2021

DOI: https://doi.org/10.22214/ijraset.2021.33397

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: 7.429
Volume 9 Issue III Mar 2021- Available at www.ijraset.com

Generating Image Descriptions using Attention Mechanism

Ms. Malge Shraddha. V[1], Prof. Shah H. P.[2]
[1, 2]*Computer Science and Engineering, M.S. Bidve Engineering College, Latur, Dr.Babasaheb Ambedkar Technological University, Lonere.*

Abstract: ***In simple terms, Image Captioning means the task of describing an image. Description comprises all the details which describe the objects within the image and their spatial connectivity. This task is performed very effectively by a human being and there are various methods to make a machine do so. In this project, we develop the encoder –decoder model to caption the image, to improve its accuracy we are using the visual attention mechanism. We used the Local/Bahdanau attention mechanism which attends to only a subset of words and is computationally simpler than global attention. We are using BLEU as a metric to evaluate the correctness of the generated caption.***

Keywords: Image captioning, Neural Network, CNN-RNN, encoder-decoder framework, Local/Bahdanau Attention, BLEU metric

I. INTRODUCTION

In recent years, the field of AI(Artificial Intelligence) is gaining so much attention due to its capabilities in all domains like computer vision, natural language processing, machine learning, neural networks, etc. Image captioning is one of the primary goals of computer vision. Image captioning is not only about identifying the objects in an image but also the relationships among the objects, like how the objects are linked with each other. That is why image captioning is viewed as a difficult problem.

The growth in this area makes AI tasks more advance than only classification and detection. The image captioning task combines working with images and text, for this purpose we need the convolutional neural network for extracting features from images, and the decoding part is handled by RNN (Recurrent Neural Network). As shown in fig. 1 The encoder-decoder framework (the classic image captioning model) encodes the image, using pre-trained CNN and generate the hidden state h. Then it uses LSTM(Long Short Term memory) to decode that hidden state h and generate each word of the caption recursively. Convolution Neural Network(ConvNet) is a deep learning algorithm that consists of an input layer, hidden layers, and an output layer. The hidden layer comprises of the layers which perform convolution on images that perform multiplication or dot product, the other hidden layers are like ReLU(Rectified Linear Unit), Pooling layer, fully connected layer, and normalization layer. A recurrent Neural Network(RNN) is used for natural language processing[12]. RNNs are developed to make use of sequential information as in sentence formation, making use of the previous word predicting the next word. LSTM RNNs are the special kind of RNN used to remember the information for a longer period.

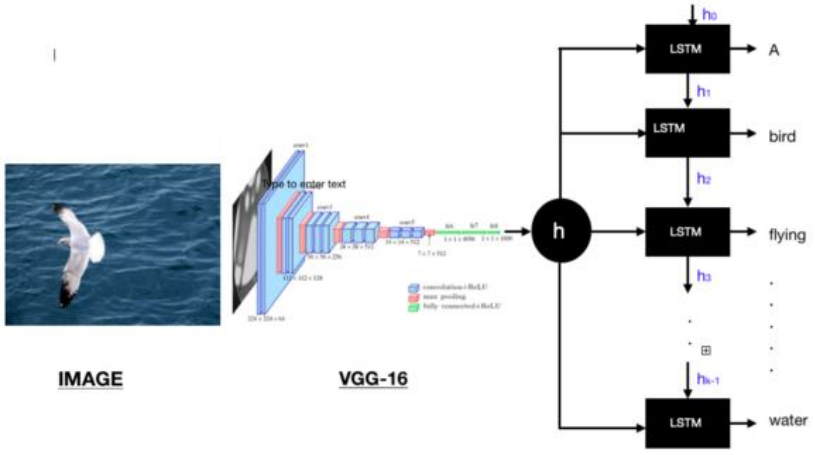


Fig. 1 A Basic Image captioning model

As the basic captioning model has some flaws. Like RNN used in the model generates a word for only the part of an image and unable to describe the essence of an entire image, this is exactly where an Attention Mechanism is helpful. With an Attention mechanism, the image is first divided into n components, and we compute *h1,h2,…hn,* the representation of each component with the help of Convolutional Neural Network(CNN). When the RNN generates a new word, the attention mechanism focuses on the important part of an image, so the decoder only uses specific parts of the image.

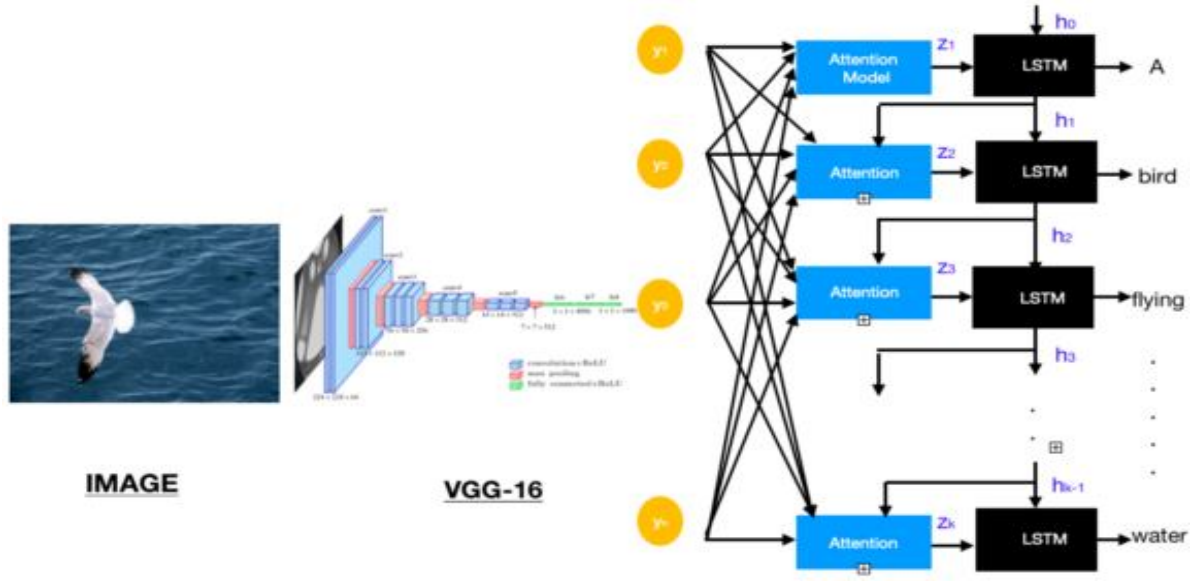


Fig 2. Image captioning with attention mechanism

II. RELATED WORK

Image captioning/ object detection is a task bridging the gap between scene understanding to generating meaningful descriptions for the corresponding image. Image captioning models can be broadly divided into two main categories: a model generating handcraft features based on statistical probability language model and a neural network model based on encoder-decoder language model which extract deep features.

A. Handcraft Features with Statistical Language Model

This model is a small system based on maximum likelihood estimation, which directly learns the visual detector and language model from the available image descriptions. Fang et al. [3] first analyze the image, detect the object, and then generate a caption. Words are detected by applying a convolutional neural network (CNN) to the image area [9] and integrating the information with MIL [10]. The structure of the sentence is then trained directly from the caption to minimize the priori assumptions about the sentence structure. Finally, it searches for most identical sentence as a caption for the image.

B. Deep Learning Features with Neural Network

The Recurrent Neural Network has gained a lot of attention in the field of deep learning. Originally RNN was used in Natural Language Processing field. RNN is used to convert text and speech to each other. Generating image descriptions based on encoder-decoder method is proposed with the application of RNN. Wherein the encoder is Convolutional Neural Network(CNN), and the features of the last i.e fully connected layer are the features of image and decoder is RNN, which is mainly used for generating image descriptions. Dealing with the sentence formation part RNN has some flaw which is it can only remember the content of the previous time unit and there comes LSTM(Long Short Term Memory), the special kind of RNN which solves the problem of gradient disappearance.

In the field of Image Captioning, Kiros et al. [2] get all the credit for the first attempt where they develop a joint multimodal embedding space providing a natural way to perform ranking and generation. To modify this version, Vinyals et al. [1] used Long Short Term Memory(LSTM) in place of a regular recurrent neural network. A. Farhadi proposed the image captioning model based on information retrieval the score generated for objects in an image is compared with other images to generate captions. On the other hand, Karpathy et al. [12] proposed learning a joint embedding space for ranking and generation.

Caption Generation is also a structured learning problem. As the input i.e image and output i.e sentence both need structural learning to be understood by a machine. The image has many features and objects to be considered and the relationships among these objects need to be captioned and there should be alignment between words in a caption with the spatial region of an image. So to address the structures properly we are using the attention mechanism in our work. Hence we have followed the Show, Attend and Tell architecture by Xu et al. [4] which generates captions for images using the attention mechanism. The attention mechanism tries to learn the latent alignments between the output words of the caption and the objects within the image. In our evaluation, we use the standard metric such as BLEU which tries to measure how accurately the caption is generated

III. IMPLEMENTATION

In this section, we tend to describe our implementation strategy for a similar basic encoder-decoder framework with the slight modification of the eye mechanism in it. As shown in Fig. 2, we tend to enclosed the eye module within the classic model of image captioning. let's have a look at however that works. If we tend to expected i words, the hidden state of LSTM is hi. we tend to choose the acceptable a part of a picture with hi as a context. Then the output of attention model zi {the illustration of the filtered image}is input to the LSTM, that successively predicts a brand new word and a brand new hidden state hi+1[13].

A. Approach

The encoder-decoder image captioning system encodes the image, employing a pre-trained Convolutional Neural Network {Here we tend useVGG16} that might turn out a hidden state. Then, it decodes this hidden state by mistreatment LSTM and generates a caption for every sequence part, outputs from previous components are used as inputs, together with new sequence knowledge. This offers the RNN networks a form of memory which may create captions additional informative and context-aware. However RNNs tend to be computationally dearly-won to coach and assess, thus in apply, memory is proscribed to simply some components. Attention models will facilitate address this downside by choosing the foremost relevant components from associate degree input image. With associate degree Attention mechanism, the image is 1st divided into n components, and that we reason a picture illustration of every once the RNN is generating a brand new word, the eye mechanism is specializing in the relevant a part of the image, that the decoder solely uses specific components of the image[14].

B. Types of Attention Mechanism

The ability of self-selection is named Attention. Attention mechanism enhances the ability of the neural network by creating it able to focus solely on the set of input to extract options. Attention is broadly classified into 2 types: one.Global Attention and local Attention. The attention-based models vary from basic encoder-decoder framework solely within the decryption half. and therefore the Global and Local mechanisms dissent from one another in an exceedingly approach they reason the context vector c(t).

1) *Global Attention (Luong's Attention):* Global attention needs more computation than local since it places the attention on all source positions of input

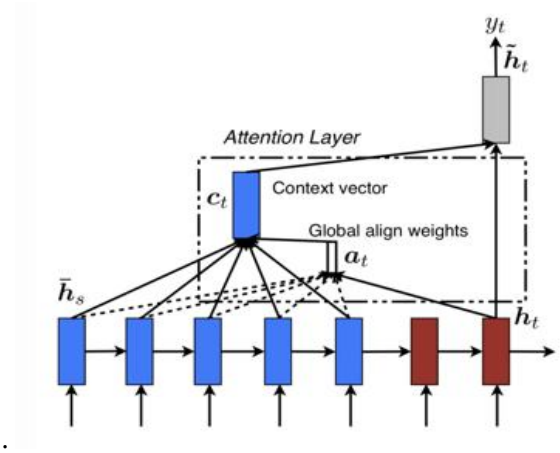


Fig. 3 Global Attention Model

In this mechanism, to derive the context vector (c(t)), all of the encoder's hidden states are taken into consideration. To calculate c(t) we compute a(t) which is the alignment vector, derived by computing the similarity between h(t) and h_bar(s) where the former is the source hidden state and later one is the target hidden state.

2) *Local Attention (Bahdanau Attention):* Global attention attends to all the source inputs making the attention mechanism computationally expensive. To beat this deficiency, local attention focuses solely on a tiny low set of hidden states of encoder per target word.

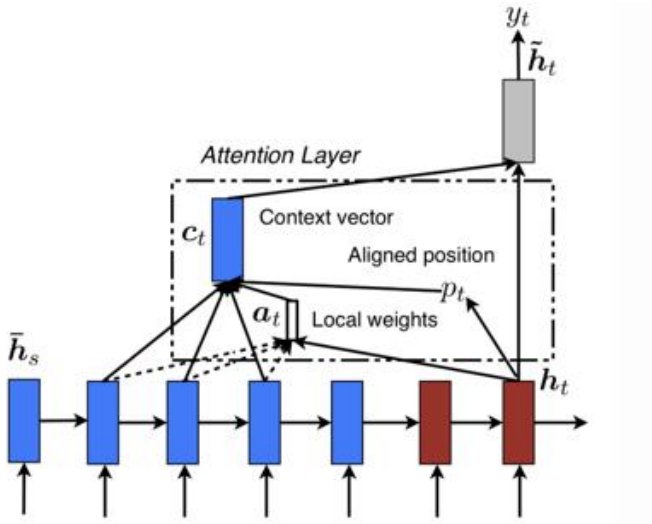


Fig. 4 Local Attention Model

C. Working of Attention Mechanism

In general, giving attention to some object within the image means we have to focus particular location of the image. Each object comes from different pixels in an image. Even though the VGG16 representations, we use do not contain any location information but the convolution layers correspond to some location of that image.

Let's say, for example, we have the following VGGNet and the output of the 5th convolution layer is 14*14 size feature map. This 14*14-pixel location corresponds to some part of an image, which means we have 196 such pixel locations, each having 512-dimensional representation. Attention is learned over these locations which in turn correspond to actual locations within images.

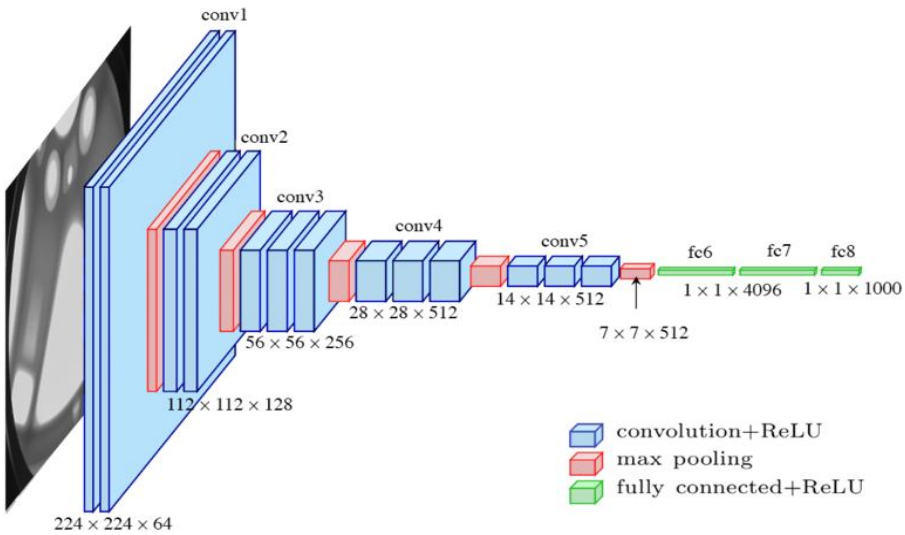


Fig. 5 VGG-16 encoder

Local attention 1st finds associate degree alignment position so calculates the attention weight within the left and right windows wherever its position is found and eventually weights the context vector. The advantage of local attention is to scale back the value of the attention mechanism calculation[7].

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue III Mar 2021- Available at www.ijraset.com

In the calculation, the local attention isn't to contemplate all the words on the linguistic communication aspect, however to predict the position of the linguistic communication finish to be aligned at this decryption per a prediction operate so navigate through the context window, considering solely the words among the window[7].
As shown in the figure below 5th convolution block is represented by 196 locations which can be passed in different time steps.

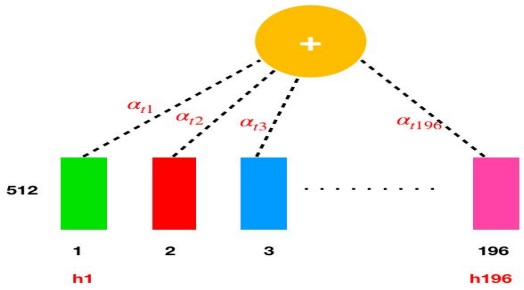


Fig. 6 5th convolution block

All hidden states of the encoder and therefore the decoder are used to generate the context vector. The local attention mechanism aligns the input and output sequences, with associate degree alignment score parameterized by a feed-forward network. It helps to listen to the foremost relevant data within the supply sequence[7]. The model predicts a target word based on the context vectors associated with the source position and the previously generated target words. Let's take a glance at all these in the form of equations:
General Score:

$$e_{jt} = f_{ATT}(s_{t-1}, h_j)$$

Where,
e_{jt} means at every t[th] timestep of DECODER, how important the j^{th} is the pixel location in the input image.
S_{t-1} is the previous state of DECODER.
h_j is the state of the ENCODER
f_{ATT} is a simple Feed Forward Neural Network which is a linear transformation of input ($U_{attn}*h_j+W_{attn}*s_t$) and then a non-linearity(tanh) on top of that and then again one more transformation(V^T_{attn}).
It is a scalar quantity

$$f_{ATT} = V^T_{attn} * tanh(U_{attn} * h_j + W_{attn} * s_t)$$

where,

$$V^T_{attn} \in R^d, U_{attn} \in R^{d*d}, W_{attn} \in R^{d*d}, s_{t-1} \in R^d, h_j \in R^d$$

To get the probability distribution we do softmax,

$$\alpha_{jt} = Softmax(e_{jt})$$

$$i.e. \quad \alpha_{jt} = \frac{e^{e_{jt}}}{\sum_{k=1}^{T_x} e^{e^{kt}}}, \text{ such that } \sum_{j=1}^{T_x} \alpha_{jt} = 1 \text{ and }, \alpha_{ij} \geq 0$$

Now, we have the input, we feed the Weighted sum combination of input to the DECODER

$$C_t = \sum_{j=1}^{T} \alpha_{jt} h_j \text{ such that } \sum_{j=1}^{T_x} \alpha_{jt} = 1 \textbf{ and, } \alpha_{ij} \geq 0$$

where, C_t is the context vector i.e weighted sum of input.

$$s_t = RNN(s_{t-1}, [e(\hat{y}_{t-1}), c_t])$$

where, s_{t-1} is the previous state of DECODER
$e(y\hat{}_{t-1})$ is a previous predicted word
and c_t *is the context vector.*

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue III Mar 2021- Available at www.ijraset.com

IV. EXPERIMENTS

In this section, we are going to discuss datasets, retrieval methodology, and the evaluation metric.

A. Dataset

For any deep learning problem, we need a dataset to train the neural network accordingly. Considering the processing capacity of our system and the availability of computing power, we have chosen the Flickr8k dataset. Flickr8k dataset provides 8k images with 5 captions per image. This 8k is divided in such a way that it allows 6k among 8k for training the network, 1k for validation, and the rest of 1k for testing purposes. The raw data is then cleaned i.e remove punctuation, single characters, numeric values, etc. Dataframe is made so that all the 5 captions are attached along with the image name. <start> and <end> tokens are tagged to the captions to make the machine understand the start and end of the caption. The images are reshaped to the same size of 224*224. The pre-trained ConvNet VGG16 is defined in a way where the softmax layer is removed since we do not want to go for image classification. The newly defined model is then trained on a training set for 20 epochs.

B. Retrieval Methodology

There are two approaches to evaluate the captions: Greedy Approach and Beam Search.
The Greedy approach is also called Maximum Likelihood Estimation(MLE). We select the word which is the most likely one according to the model. The word with maximum probability is selected. In the Beam search technique, top k predictions are chosen, feed them again in the model and then sort them using the probabilities returned by the model. So, the list will always contain the top k predictions. In the end, select the one with the highest probability and go through it till we encounter <end> or reach the maximum caption length.
We are using the greedy one among the two.

C. Evaluation Metric

We are using the BLEU[5] score as a metric to estimate the accuracy of the generated Caption for the test set. The BLEU is simply taking the fraction of n-grams in the predicted sentence that appears in the actual caption. BLEU is a well-known metric used to measure the similarity between one hypothesis sentence to multiple reference sentences. It returns the value between 0 and 1as a BLEU score. A score near 1 means the two are very much similar.

V. RESULTS

This section consists of the outcomes of our image captioning project with an attention mechanism. Our model generates captions for the test dataset. As we can see, our model performs quite well as it generates captions close to the real one. Sometimes it predicts wrong captions(look at image No.8), and sometimes it outperforms and generates a better caption than the real one like in image No. 6

	BLEU score: 75.59289460184544 Real Caption: two brown dogs run through the grass Predicted Caption: two dogs run through the grass
	BLEU score: 33.40135926488845 Real Caption: little boy in blue doing bike jumps Predicted Caption: boy on bike

	BLEU score: 18.502935537050632 Real Caption: man drives an atv down dirt road with power line behind him Predicted Caption: man on snowmobile in rural area
	BLEU score: 16.5748386032949 Real Caption: the wet dog has retrieved the pink purse with aqua handles Predicted Caption: black dog is running in the grass with pink purse in its mouth
	BLEU score: 47.14045207910317 Real Caption: men play in soccer game Predicted Caption: two men are playing soccer on the field
	BLEU score: 0 Real Caption: dog bites an object <unk> by person Predicted Caption: golden retriever is digging in the air to bite large toy
	BLEU score: 55.06953149031837 Real Caption: man rides wave on surfboard Predicted Caption: man rides surfboard

	BLEU score: 24.618298195866547 Real Caption: small dog leaps across brown pillow Predicted Caption: white dog with brown and black and black and black and black and black and black and black and black and black and black and black and black and black and black
	BLEU score: 45.78227398676668 Real Caption: people running away from seagulls on the beach Predicted Caption: two people hopping hot sand
	BLEU score: 23.505403213046534 Real Caption: the cricketer is swinging bat in the air Predicted Caption: man swinging bat

A. Comparision

Let's Compare our attention-based model with Karpathy et al.[] by trying them on the MS COCO dataset. We have shown both the descriptions. There are two captions/descriptions generated: (A). specifies caption generated by NeuralTalk and (B) specifies the caption by our model, the attention-based model. When we look at the results, we can say that model generates incorrect or irrelevant captions for the images which are not semantically identical to the training set. While talking about comparison, we sometimes perform better than Neuraltalk. Like in images [8] and [9], our model generated the caption that is partially correct than the NeuralTalk which described it incorrectly. And looking at the image[6], when we come across the random unknown features of the image, we cannot describe it correctly.

		
B. baseball player is throwing ball in game. *C.* : baseball pitcher with the ball [1]	A.young girl in pink shirt is swinging on swing B. girl in blue shirt gestures with branches [2]	A. black and white dog jumps over bar. B. black and white dog is jumping over the ground with its owner [3]
		
D. man in blue wetsuit is surfing on wave. *E.* man in blue bodysuit boarding blue waves [4]	A. two young girls are playing with lego toy. B. while two children play with neon colors baby is looking in the camera [5]	A. woman is holding bunch of bananas. B. some young children are gathered outside [6]
		
A. a group of young men playing a game of soccer B. group of children playing soccer [7]	A. a young boy holding a frisbee in his hands. B. two young girl in yellow dress is swinging on the ground eating [8]	A. man is holding baseball bat in his hands B. young asian girl with white shirt is kicking soccer ball [9]

VI. CONCLUSIONS

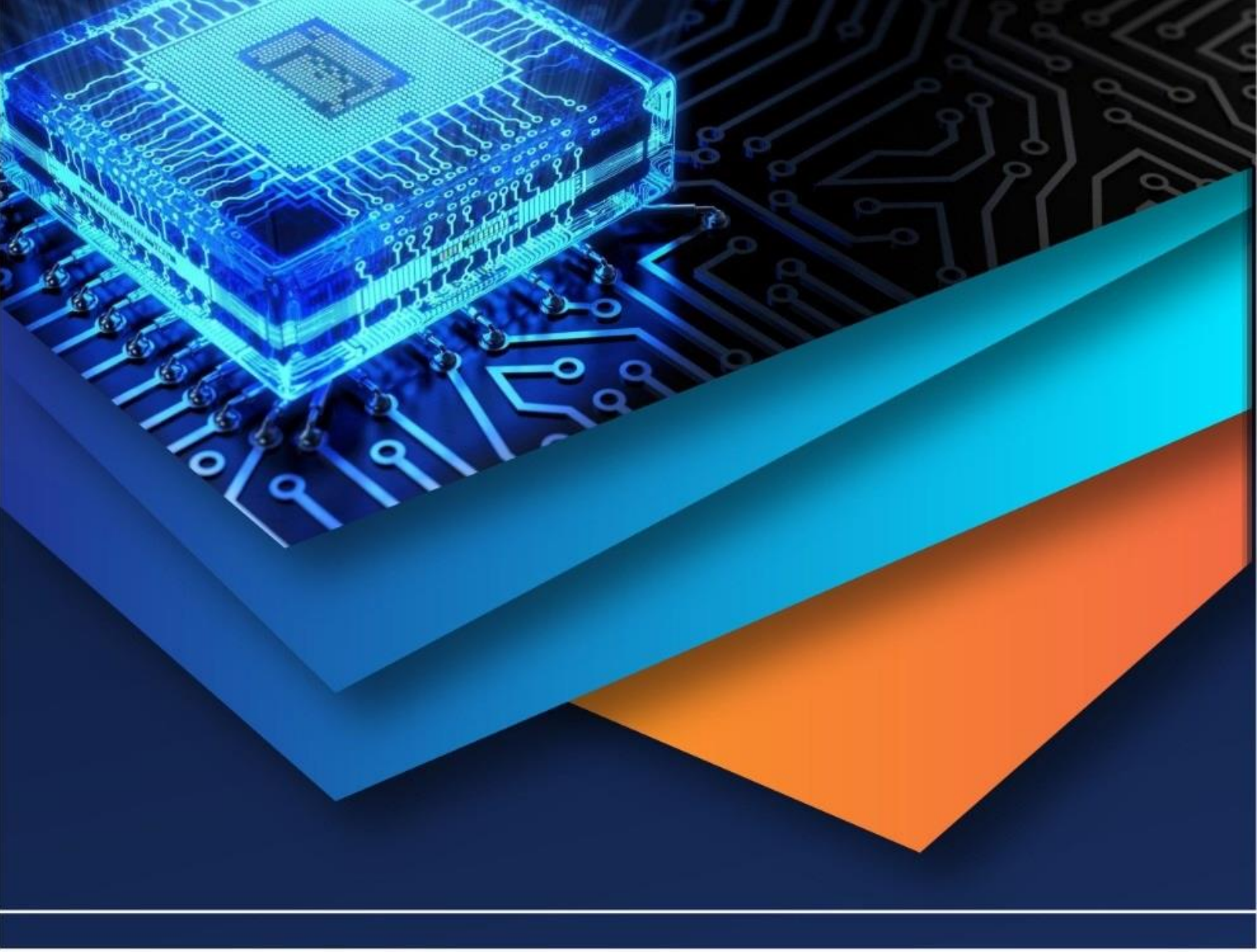
To conclude the overall performance for this image captioning model, I would say adding the attention mechanism improves the correctness of generated caption. Sometimes the predicted one is more related to the image than the real one. And another thing is that we must remember that the images in the training set and the testing set must be semantically related, for example if it's not the case then no machine learning model could perform better. When there is an AI task there is always the future scope. A lot of modifications can be made to carry out much better results than this: using the larger dataset, changing the model architecture(adding Batch Normalization layer, dropouts, etc), implement different attention mechanisms.

VII. ACKNOWLEDGMENT

I would like to express my sincere gratitude to Prof. Shah H. P. for her relevant guidance throughout my research and I would like to thank her for the support and inspiration she gave me during this whole phase. I would also like to thank our respected head of the department Prof. Tandle S. R. and all the faculty members and friends for their cooperation. And the special thank you to Google for their awesome products like colab and drive. Google Colab provided the free GPU for deep learning and storage is provided by Google Drive.

REFERENCES

[1] Vinyals O, Toshev A, Bengio S, Erhan D, Show and tell: a neural image caption generator. CVPR., 2015
[2] Kiros R, Salahutdinov R, Zemel R ,Multimodal neural language models. ICLR, pp 595–603,2014
[3] H. Fang, S. Gupta, F. Iandola et al., “From captions to visual concepts and back,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, June 2015.
[4] Kelvin Xu,Jimmy Lei Ba, Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, 2016.
[5] Kishore Papineni, et al. “BLEU: a Method for Automatic Evaluation of Machine Translation“,2002.
[6] https://www.tensorflow.org/tutorials/text/image_captioning
[7] Haoran Wang , Yue Zhang, and Xiaosheng Yu (2020), An Overview of Image Caption Generation Methods,2020
[8] Biswas, R., Barz, M. & Sonntag, D. Towards Explanatory Interactive Image Captioning Using Top-Down and Bottom-Up Features, Beam Search and Re-ranking. Künstl Intell **34,** 571–584 (2020). https://doi.org/10.1007/s13218-020-00679-2
[9] R.Girshick, J. Donahue, D. Trevor, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587, Columbus, OH, USA, June 2014.
[10] C. Zhang, J. C. Platt, and V. Paul, “Multiple instance boosting for object detection,” in Advances in Neural Information Processing Systems 18, pp. 1417–1424, MIT Press, London, UK, 2005.
[11] Andrej Karpathy, Li Fei-Fei, Deep Visual-Semantic Alignments for Generating Image Descriptionshttps://cs.stanford.edu/people/karpathy/cvpr2015.pdf.
[12] Zhihao Fan, Zhongyu Wei, A Question Type Driven Framework to Diversify Visual Question Generation
[13] https://medium.com/heuritech/attention-mechanism-5aba9a2d4727
[14] https://www.analyticsvidhya.com/blog/2020/11/attention-mechanism-for-caption-generation/



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY