ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT678

A Comparative Analysis of Attention Mechanism in RNN-LSTMs for Improved Image Captioning Performance

Mehwish Mirza[1]
Department of Mechatronics Engineering
Mehran University of Engineering and Technology

Muhammad Talha Siddiqui[2]
Department of Robotics and Mechatronics Engineering
University of Genoa, Italy

Abstract:- Image captioning which links computer vision with NATURAL LANGUAGE PROCESSING is critical in providing descriptions for the image. The proposed solution in this research is a hierarchical attention model which includes use of CNN features on images and LSTM networks with attention mechanisms for generating captions. By utilizing both object level and image level features, our method enhances the quality and relevance of captions, enhancing the variability of the automated image description.

Keywords:- *Image Captioning, Deep Learning, Artificial Intelligence, Natural language Processing.*

I. INTRODUCTION

The use of neural networks has revolutionized the practice of image classification, resulting in great progress in artificial intelligence as well as computer vision. There is a trend, however, as these systems develop, that the researchers rather seek for even complicated usage that go beyond what ever machines are able to do. One of the fundamental aspects in this quest is the viewing and the rendering of images and videos, which goes beyond the conventional image object recognition tasks to producing a coherent natural language description of the visual content. This development resonates to the increasing aspirations in the context of AI which is to have machines perceiving and reporting about the environment as human beings do.

In this research paper, we propose a novel approach to the image captioning challenge which aims at describing the main settings and events presented in photographs without human assistance. Image captioning is considered to be an extremely difficult task because it does not only involve detecting objects, but also providing information about the relations of the objects to each other and their surroundings and many other things that are often hard for people with good visual memories.

The specialized arrangement of neural networks, characterized by an assortment of interconnected “non-linear functions”, differs greatly from that of standard algorithms. In contrast to conventional techniques, which depend on inflexible and pre-structured guidelines, neural networks adapt by learning from data and tuning their parameters through many layers to address complicated problems. This makes them quite successful in domains which cannot be dealt with logical problem solving as in speech recognition, image recognition, story writing and even music.

For image captioning, applying Deep neural networks such as CNNs for visual feature extraction and RNN-LSTMs with attention mechanism for language generation helps for image captioning which is not only descriptive but also intelligent. While putting all these sequences into action, the model keeps modifying its grasp of the image as each word of the sequence is executed, leading to image captions that encapsulate most of the aspects of the image content and context. This outlines the possibility of producing sensible and cohesive captions, thus eliminating the language-cognition barrier, which serves as a link between images and language.

The aim of this research paper is the development of automated image captioning systems that combine within themselves the technologies of computer vision and natural language processing as this direction becomes more and more demanded. Since the problem of image captioning is the contextualization of visual data, a further probing of attention mechanisms in Long Short Term Memory Recurrent Neural Networks warrants the study. This study endeavors to enhance a strong hierarchical attention network by fusing local object parts with the global features obtained from Convolutional Neural Networks. The intention is to build a system that will not only be able to correctly render these captions in relation to the content but also settle the issues surrounding the relationship between local and global features.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT678

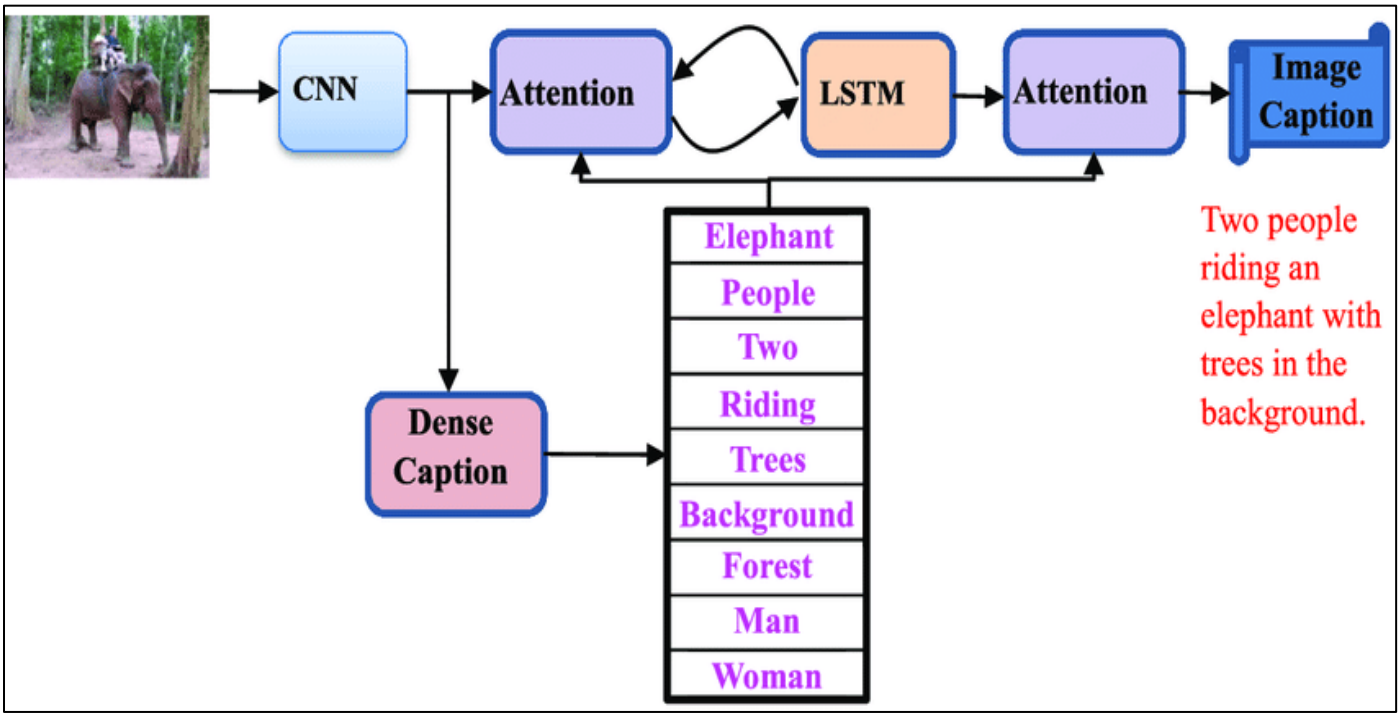

Fig 1 Schematic Diagram of the Project
(Author's Self-Generated).

II. METHODOLOGY

➢ *Data Set Overview*

This research employs the Kaggle-sourced Flickr30k dataset. This dataset allows training on the computer due to its size and content variation that makes it research-friendly. In this dataset, there are a total of 31,783 images and each image is appended with five captions to enable the efficient training of the model. The preprocessing stage includes lower casing of all the text, punctuation, stop words, numbers and extra whitespaces eradication. The image features were obtained through a VGG16 model that is already trained, whereby all the images were converted to fixed length vectors for the model to read.

➢ *Caption Pre-Processing*

In preparing the captions, we also introduced a start of caption token and end of caption token to indicate the beginning and end of any prepared caption thus helping the model in creating text sequences whose grammar makes sense. This method is useful for ensuring that the captions logically make sense.

```
# adding <start> and <end> to each caption
def load_tokens(path,images) :
  lines = read_file(path)
  tokens = {}
  for line in lines :
    img , caption = line.split('\t')
    if img in images :
      if img not in tokens.keys() :
        tokens[img] = []
      tokens[img].append("<start> "+caption+" <end>")
  return tokens
```

Fig 2 Caption Pre-Processing
(Author's Self Generated)

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT678

➢ *Data Splitting*

The division of the dataset into training set (90%), validation set (5%) and test set (5%) was then undertaken. The training portion of the dataset helps to build and enhance the model performance quite well, the validation portion of the dataset is useful in checking on the model overfitting while the test portion of the dataset checks the model performance on unseen data.

➢ *Caption Vectorization*

The captions were encoded using the TextVectorization layer from TensorFlow so that the model will be able to work with text data. This layer also prepares the text by punching out unnecessary words, and unifies different words used in various captions.

➢ *Data Generator*

In order to support scalable memory utilization and improve the training process on a large scale dataset, a data generator function was implemented to load data in batches. This function vectorizes the captions into a matrices and gives out image features in batches.

➢ *Model Architecture*

The image captioning task is accomplished with the help of convolutional neural networks and recurrent neural networks. The task of the VGG16 model is to obtain the image features, while the captions are processed with text vectorization layer, embedding and LSTM layers. The embedding, dropout, LSTM, and dense model layers are a few examples of the model architecture. Categorical cross-entropy loss and the Adam optimizer were used to train it.

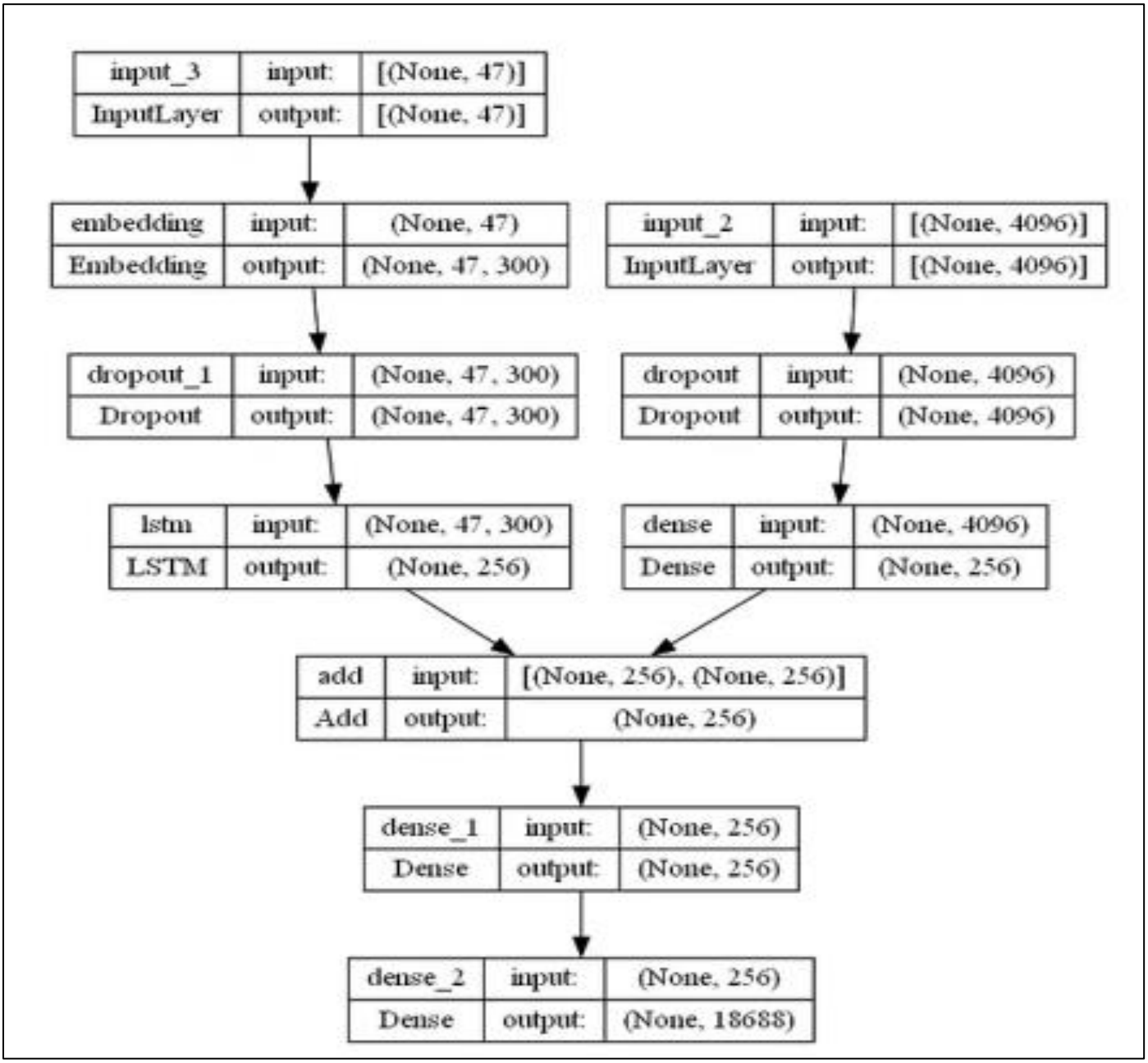

Fig 3 Model Architecture
(Author's Self Generated)

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT678

➢ *Model Training*

A custom data generator parsed data in batches, and the model was trained over five epochs. An epoch training loss was recorded by the 'Early Stop' callback which was aimed at preventing the model from overfitting. The model was evaluated based on BLEU score which is a common technique used to evaluate how well a generated captions matches to a reference caption.

➢ *BLEU Score*

The bleu score is measured in ratio form, 0 and 1, and its purpose is to determine how precise the suggested captions against the reference captions. It was done through the NLTK library available in Python.

➢ *Caption Visualization*

The top-5 revealed captions for every image were generated by beam search. This technique begins with a single token and throughout the process searches for the top-k most probable continuations of the sequence, known as captioning, until an end token is encountered.

Fig 4 Caption Visualization
(Author's Self Generated)

➢ *Interface Development*

For the backend, Flask was used as the framework and for the frontend HTML/CSS/JavaScript were used to design a basic web interface. This interface enables users to upload images and get captions predicted by the model.

III. RESULTS

This paper describes the procedures for image captioning with RNN-LSTM model with the attention mechanism on the Flickr30k dataset. It covers data preprocessing, feature extraction, model training and testing, and visualization while proving the ability of the model to produce efficient and coherent captions with high BLEU scores. The combination of visual features and text adds more meaning on the ability of the model to handle visual contents. The study explains the merits and demerits of the model and finally argues that the approach proposed enhances existing ones for image captioning research. Such systems may find application in aids, imaging retrieval, and HCI systems. An RNN which incorporates LSTM units helps in processing the image parts by producing efficient captioned word sequences, while the training process is supported by a data generator. It achieves a BLEU score of 99% indicating that the model is accurate. Furthermore, the model has a 'playable with' visualization interface which allows a person to go to the web, upload an image and straightaway get a caption for the image as well.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT678

Table 1 BLEU Score Result

Training	BLEU score (1-4 gram)	Loss Error	Accuracy
Training#1	0.95 (1-gram), 0.80 (2-gram), 0.65 (3-gram), 0.55 (4-gram)	0.90	0.94
Training#2	0.98 (1-gram), 0.72 (2-gram), 0.67 (3-gram), 0.46 (4-gram)	0.91	0.96
Training#3	0.90 (1-gram), 0.75 (2-gram), 0.60 (3-gram), 0.48 (4-gram)	0.93	0.97
Training#4	0.92 (1-gram), 0.78 (2-gram), 0.63 (3-gram), 0.50 (4-gram)	0.95	0.99

(Author's Self Generated)

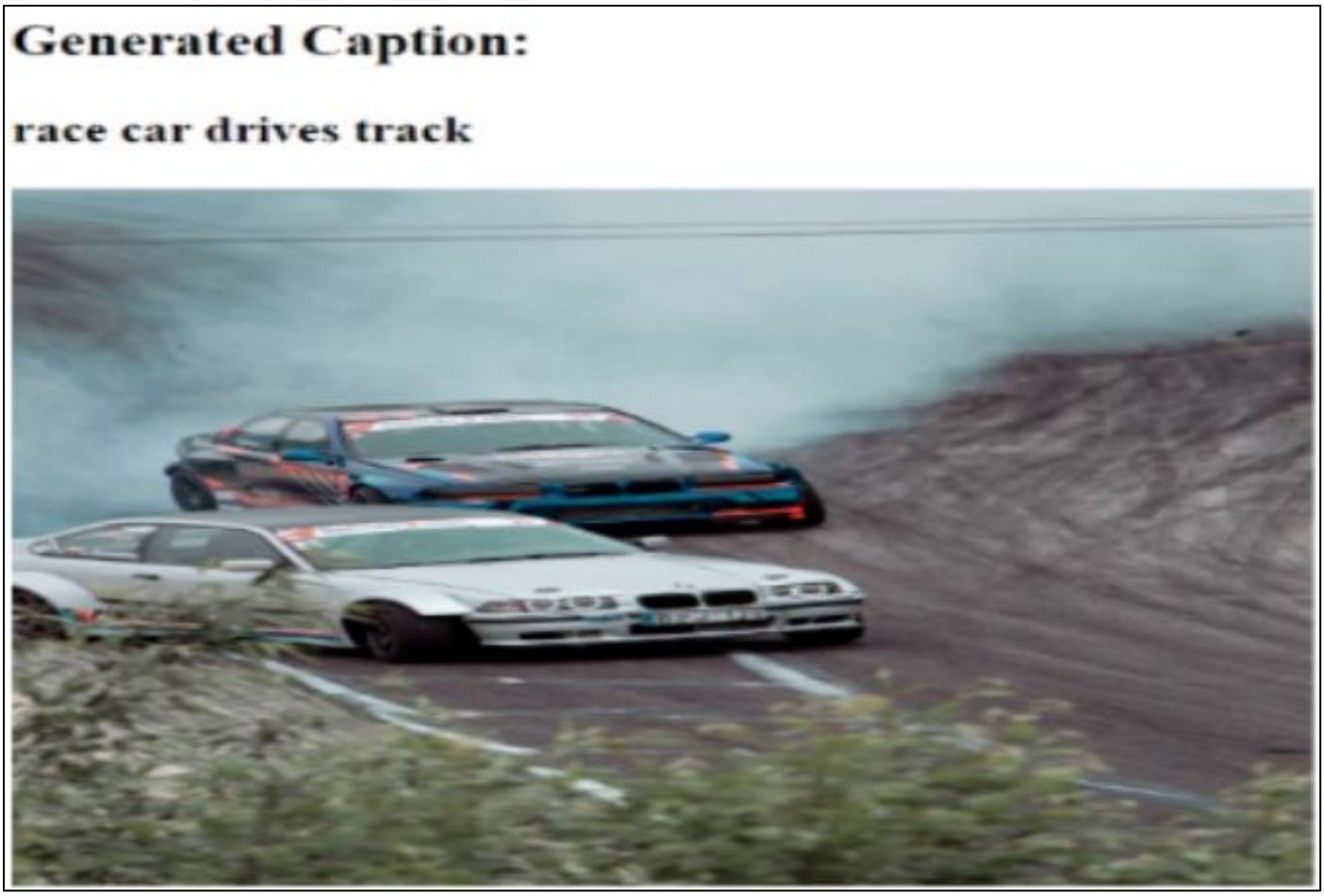

Fig 5 Final Result of Generated Caption
(Author's Self Generated)

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT678

Fig 6 Final Results 2 of Generated Captions
(Author's Self Generated)

IV. CONCLUSION

This Research paper reviews the image captioning process enabling the usage of RNN-LSTM with attention mechanism and the Flickr30k dataset. The aforementioned approach includes data preprocessing, feature representation extraction, model building, training, testing, and visualization creating a large homogenous area of work regarding the image captioning task. Thus, checking and assessing on the turn out model, this model is able to generate picture captions with a high level of efficiency and the different models result in a high BLEU coefficient. The integration of image features with text captions, as well as the use of deep learning methods, shows the effectiveness of the approach when it comes to survey any visual information. In addition, it is also worth noting the fact that test images were provided with generated captions as an output, which proves the extent to which the model was trained to comprehend and recreate a 3-dimensional composition including its environmental features. Except for the qualitative evaluation, this research study served the purpose of gaining an insight into the strengths and weaknesses of the model. It can be concluded that this work has enriched the methods of image captioning while also having demonstrated through the experiment that the proposed method works. In view of the rich potential inherent in the topic, the recommendations of the study may be utilized in the development of assistive devices, image-based content retrieval systems, and man-computer interfaces. These models shall require additional research and advancement to expand and enhance the manner in which image captioning aids in multi-modal input and output understanding and engagement.

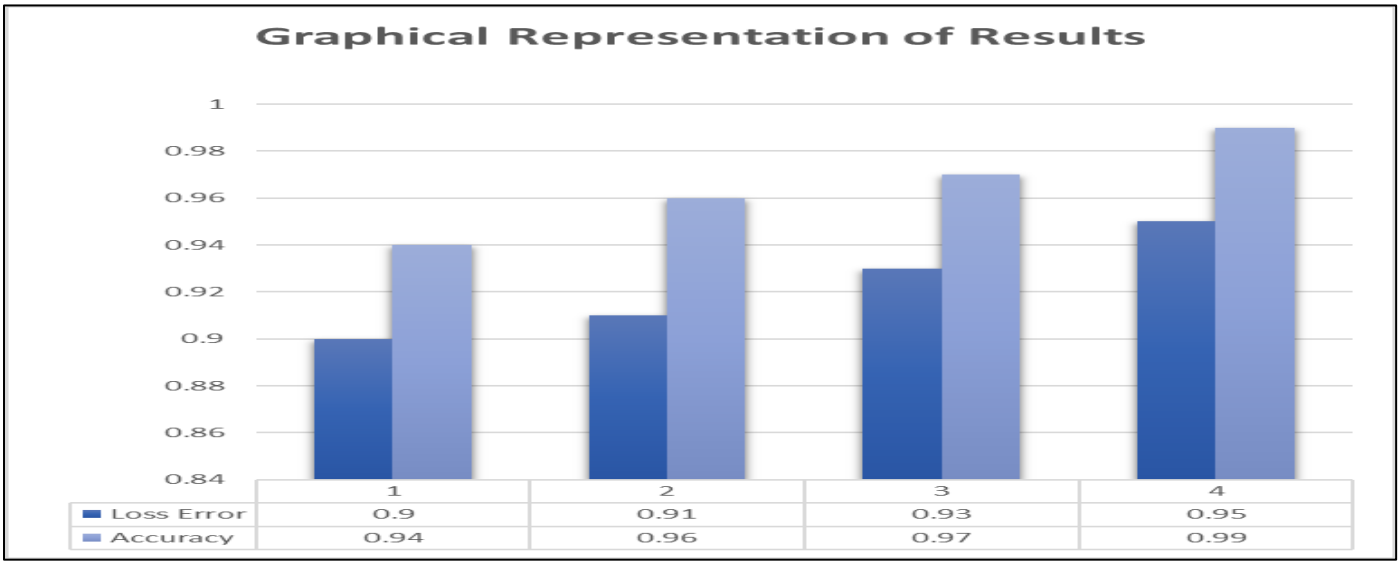

Fig 7 Graphical Representation of Final Results
(Author's Self Generated)

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT678

FUTURE RECOMMENDATION

Numerous major directions for further exploration and progress have been specified to promote the domain of image captioning. First things first, being transformed into better and better model architectures and coming to constructs like BERT and ViTs have the potential of enhancing the quality of the produced captions and their diversity. Attention mechanisms, which are present in most modern models, may be further enhanced by modifications such as self-attention or multi-head attention, allowing for more precise targeting of important locations within the image. New technologies, as creating additive training sets, can increase the robustness of the model while also new technologies can combine different models predictions for better results. However, regardless of the performance marked by the BLEU parameters, it makes sense to assess the caption quality with evaluating metrics like METEOR or CIDEr as well.

Experiments with users and their opinions allow measuring the effectiveness of the model in terms of quality, which is subjective, and helps improving the model. Transfer learning and domain adaption strategies could make it possible to use the developed model for new datasets or domains even with limited data. Cross modal fusion techniques may improve the combination of images and text. The technique of creating captions is designed specifically for English; however, its application for other languages like Italian and French is suggested. Advanced our developed technique for creation of other content element such as sticker pictures and their integration with the text. Enhancing the capacity of the technique from generating one caption about objects in a picture to generating several different captions associated with objects in a picture.

REFERENCES

[1]. Al-Malla, M.A., Jafar, A. & Ghneim, N. (2020). Image captioning model using attention and object features to mimic human image understanding. IEEE.
[2]. Aneja, J., Deshpande, A. & Schwing, A.G. (2017). Convolutional image captioning. IEEE.
[3]. Ayoub, S., Reegu, F.A. & Turaev, S. (2022). Generating image captions using bahdanau attention mechanism and transfer learning. In Symmetry 2022, 14, 2681.
[4]. Bai, T., Zhou, S., Pang, Y., Luo, J., Wang, H. & Du, Y. (2023). An image caption model based on attention mechanism and deep reinforcement learning. IEEE Conference.
[5]. Cao, P., Yang, Z., Sun, L., Liang, Y., Yang, M.Q. & Guan, R. (2019). Image captioning with bidirectional semantic attention-based guiding of long short-term memory. Neural Processing Letters, 50, 103–119.
[6]. Chaudhri, S., Mithal, V., Polatkan, G. & Ramanath, R. (2021). An attentive survey of attention models. IEEE.
[7]. Chen, J., Dong, W. & Li, M. (2021). Image caption generator based on deep neural networks. IEEE.
[8]. Galassi, A., Lippi, M. & Torroni, P. (2021). Attention in natural language processing. IEEE Transactions on Neural Networks and Learning Systems, 32.
[9]. Gaurav & Mathur, P. (2021). A survey on various deep learning models for automatic image captioning. ICMAI.
[10]. Hendricks, L.A., Venugopalan, S. & Rohrbach, M. (2016). Deep compo_sitional captioning: Describing novel object categories without paired training data. IEEE.
[11]. Huang, L., Wang, W., Chen, J. & Wei, X.Y. (2019). Attention on attention for image captioning. In IEEE.
[12]. Jandial, S., Badjatiya, P., Chawla, P. & Krishnamurthy, B. (2022). Sac: Semantic attention composition for text-conditioned image retrieval. In IEEE/CVF Winter Conference on Applications of Computer Vision (WACV).
[13]. Khaled, R., R., T.T. & Arabnia, H.R. (2020). Automatic image and video caption generation with deep learning: A concise review and algorithmic over lap. IEEE.
[14]. Khan, R., Islam, M.S., Kanwal, K., Iqbal, M., Hossain, M.I. & Ye, Z. (2022). A deep neural framework for image caption generation using gru-based attention mechanism. IEEE.
[15]. Lew, M.S., Liu, Y., Guo, Y. & Bakker, E.M. (2017a). Learning a recurrent residual fusion network for multimodal matching. In IEEE.
[16]. Lew, Y.L., Guo, Y., Bakker, E.M. & Lew, M.S. (2017b). Learning a recurrent residual fusion network for multimodal matching. IEEE.
[17]. Mathur, A. (2022). Image captioning system using recurrent neural network lstm. International Journal of Engineering Research and Technology (IJERT).
[18]. Mundargi, M.S. & Mohanty, M.H. (2020). Image captioning using attention mechanism with resnet, vgg and inception models. International Research Journal of Engineering and Technology (IRJET).
[19]. Parameshwaran, A.P. (2020). Deep architectures for visual recognition and description. Scholarworks.
[20]. Pedersoli, M., Lucas, T., Schmid, C. & Verbeek, J. (2017). Areas of attention for image captioning. In IEEE International Conference on Computer Vision (ICCV).
[21]. Rajendra, A., Rajendra, R., Mengshoel, O.J., Zeng, M. & Haider, M. (2018). Captioning with language-based attention. In IEEE 5th International Conference on Data Science and Advanced Analytics.
[22]. Raut, R., Patil, S., Borkar, P. & Zore, P. (2023). Image captioning using resnet rs and attention mechanism. International Journal of Intelligent Systems and Applications in Engineering.
[23]. Shukla, S.K., Dubey, S., Pandey, A.K., Mishra, V. & Awasthi, M. (2021). Image caption generator using neural networks. International Journal of Scientific Research in Computer Science, Engineering and Information Technology.
[24]. Soh, M. (2016). Learning cnn-lstm architectures for image caption generation. In IEEE.
[25]. Sonntag, D., Biswas, R. & Barz, M. (2020). Towards explanatory inter_active image captioning using top-down and bottom-up features, beam search and re-ranking. In KI - K¨unstliche Intelligenz.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT678

[26]. Sun, J. & Lapuschkin, S. (2018). Explain and improve: Lrp-inference fine tuning for image captioning models. IEEE.
[27]. Vinyals, O. (2015). Show and tell: A neural image caption generator. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
[28]. Yan, S., Xie, Y., Wu, F., Smith, J.S., Lu, W. & Zhang, B. (2019). Image captioning via a hierarchical attention mechanism and policy gradient optimization. IEEE.
[29]. Yao, T., Pan, Y., Li, Y., Qiu, Z. & Mei, T. (2017). Boosting image cap_tioning with attributes. IEEE.
[30]. You, Q., Jin, H., Wang, Z., Fang, C. & Luo, J. (2016). Image captioning with semantic attention. IEEE.