ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol 19, Issue 2, 2025

A Hybrid Model for Image Caption Generation Based on Deep Learning That Works Well

[1] Ms.Zeba Unissa, [2] Mir Faizan Ali, [3] Syed Mateen, [4] Mohd Haseeb Mohiuddin
[1] Assistant Professor,Department of CSE-AIML, Lords Institute of Engineering & Technology.
[234] Student Department of CSE-AIML, Lords Institute of Engineering & Technology.

Abstract—

With the proliferation of social media in recent years, picture captioning has emerged as a powerful tool for automatically translating full-image images into natural-language captions. In our digital culture, image captioning is crucial. picture captioning involves the use of artificial intelligence algorithms to automatically generate a textual description of a picture in a natural language. The core of the picture processing system is computer vision and NLP. Computer vision includes the use of Convolutional Neural Networks (CNNs) for object recognition and feature extraction, while Natural Language Processing (NLP) methods aid in the generation of picture captions. Object identification, position, and the semantic linkages between them in a human-understandable language like English make it a difficult challenge for machines to provide appropriate picture descriptions. Using VGG16, ResNet50, and YOLO, we want to create a hybrid picture captioning method that relies on encoder-decoders in this article. Pre-trained feature extraction models, such as VGG16 and ResNet50, are trained using millions of photos. In order to identify objects in real time, YOLO is used. It begins by merging the output of feature extraction using VGG16, ResNet50, and YOLO into a single file. Finally, the picture is described textually using LSTM and BiGRU. The BLEU, METEOR, and RUGE scores are used to assess the proposed model. A.I., LSTM, CNN, and YOLO are some of the keywords.

I. INTRODUCTION

Every single one of us encounters a vast array of real-world pictures on a daily basis in this web environment, and each of us uses our own unique set of experiences and insights to make sense of them. Although humans have the innate ability to translate scenes from the real world into words, machines have significant challenges in this area and are not nearly as efficient as humans. Because machines need human interaction and are configured appropriately, human-generated captions are still seen to be superior. Computers can now manage the complexities of picture captioning, such as object and attribute recognition, feature extraction, and the generation of syntactic and semantic captions, thanks to advancements in deep learning-based approaches [1]. The world has changed in an unexpected manner due to the revolutionary new ideas brought forth by AI's progress in image processing. Since it offers a superior platform for human-computer interaction, the picture captioning approach (Fig. 1) has broader practical applications. Image captioning has recently attracted the attention of academics and researchers as a result of its potential new uses in image processing. The two dogs in Figure 2 are likely playing with a toy, but the captions "two dogs fetching a floating toy from the ocean" or "two dogs racing through the water with a rope in their mouths" might also work. Machines couldn't match the accuracy with which our highly trained brains could characterize images. Therefore, the primary goal of picture captioning is to use deep learning to detect objects and their relationships in the image, then use natural language processing to generate a textual description, and finally, use various performance matrices to assess how well the textual description was generated. Computer vision tasks like object identification and segmentation are accomplished with the aid of well-known convolutional neural networks (CNNs) and recurrent neural networks (RNNs) and long short-term memories (LSTMs). In order to answer queries like "what," "where," and "how" pertaining to the things in a picture, CNN is able to analyze the scene and its objects.

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

Fig. 1. Image captioning.

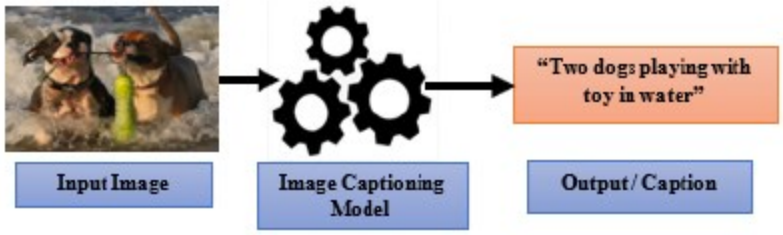

Fig. 2. Working of image captioning.

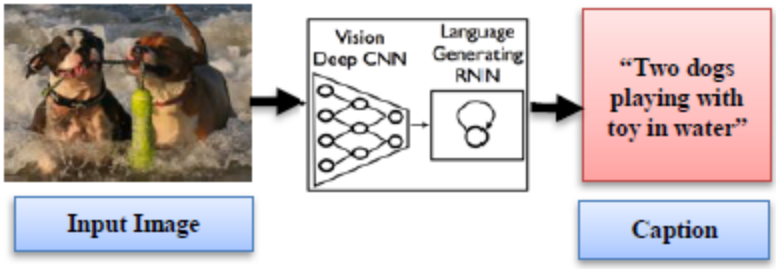

Fig. 3. Image captioning architecture.

As an example, in Figure 3, CNN is able to recognize the "dog," "toy," and "water" objects and establish a connection between them. In addition, RNN transform the shape into text by making use of the keywords that CNN outlined, taking them into consideration as a set of words. One such architecture is the encoder-decoder. Computer vision's object identification subfield makes use of a number of techniques, including YOLO, R-CNN, Mask R-CNN, MobileNet, and SqueezeDet, among others, to accurately identify visual elements.

II. LITERATURE SURVEY

The literature review on picture captioning is provided in this part. In order to create captions that seem human, many state-of-the-art methods and models have been released in the last several years. The techniques used for image captioning may be roughly categorized into three groups: template-based, retrieval-based, and encoder-decoder. An method to content selection for picture description based on geometric, conceptual, and visual aspects of the image is suggested in article [31]. Every one of these models is CNN-based; they all start with picture encoding and feature extraction, and then utilize RNN or LSTM to generate captions. By combining picture captioning with probabilistic distributions of successor and predecessor terms, researchers in article [1] created a model for image captioning. One well-known method in picture captioning is the attention and visual oriented approach. The writers in [2,3] use the attention mechanism to construct the captions. Many of the publications relied on preexisting models. Some examples of these models are the well-known encoder or CNN model Unet [13], the

Inception V3 [9-10] and VGG16 papers [1], [3-7], AlexNet [5], [7], ResNet [4-5], [12], and AlexNet [7]. In order to generate or decode picture captions, RNN [16], BiLSTM [7], and LSTM [8-10] and [15] are all viable options. Captions for images may also be created in a number of languages, including but not limited to German, Punjabi, Chinese, Japanese, and Hindi. The input picture is identified using a template-based technique, which makes use of preset objects, actions, and attributes. In order to anticipate the image's caption, the writers [18] make use of visual components such as object, action, and setting. Using a Conditional Random Field (CRF) based approach, the author of [19] extracts picture characteristics. The BLUE and ROUGE scores on the PASCAL dataset were used to test the proposed model. Since it relies on a pre-made template, it can't create captions for images of varying durations. By matching the image's attributes with existing datasets, a retrieval-based method may automatically create captions. Input images are searched for captions using comparable attributes found in the dataset. In [22], the authors provide a model for query picture feature extraction by dataset searching; in [32], they suggest a technique for caption extraction using density estimation. To generate captions for images, the authors of [25] used both visual and semantic characteristics. Each picture in the original dataset has five captions, and we want to use this dataset to train a certain model. Once the model has mastered the art of feature extraction during training, there are a number of readymade image classification models at your disposal that use cutting-edge algorithms to effectively categorize thousands of unique objects and photos. Similar to ResNet, these models provide more accurate results when it comes to picture rate categorization. These are a breeze to put into action. Machine translation and deep neural network–based picture caption synthesis both make heavy use of encoder-decoder based approaches. Both the NIC (Neural Image Caption) model and a dual graph convolution network based on encoder-decoder architecture are presented in [33] and [27], respectively. The model is straightforward; it employs convolutional neural networks (CNNs) as encoders and LSTMs and RNNs as decoders to generate picture captions.

III. RESEARCH METHODOLOGY

In this case, convolutional neural networks (CNNs) with pooling and fully connected layers are used as encoders to extract visual features from the picture. In the past, AlexNet was the go-to for compute vision problems. However, these days, transfer learning is all the rage, and there are a plethora of pre-trained CNN based models available, such as VGGNet, Inception V3, DenseNet, ResNet, etc., each with its own unique set of convolutional neural layers that can be utilized to cut down on training time. The encoder provides the data that the decoder uses to generate the final captions. The three most popular decoders are GRU, LSTM, and RNN. Long short word sequences are more suited to LSTM, whereas short word sequences work well with RNN. The suggested hybrid research approach is shown in this section. In developing this concept, we primarily set out to increase the Meteor value. The foundation of our model is the idea of transfer learning, which is rooted in an encoder-decoder technique. In this first step, we use VGG16, ResNet50, and YOLO (You Only Look Once) individually to extract picture characteristics. In contrast to VGG16, an object identification and classification method pretrained on the ImageNet dataset, YOLO is an effective real-time object detection technique created in 2015 by Joseph Redmon et al. With its sixteen convolutional layers, this design is a deep CNN. A deep convolutional neural network (CNN) with 50 convolutional layers, ResNet50 can categorize over a thousand different types of objects. The second step is to combine the picture characteristics obtained from VGG16, ResNet50, and YOLO, and to remove any duplicate words. The final step involves utilizing the BiGRU and LSTM to produce captions. Natural language processing makes use of a Neural Network design known as BiGRU. Both forward and backward inputs are handled by two GRUs in this design. LSTM, short for "long short-term memory," is an architecture for recurrent neural networks that can recognize relationships between objects via the use of feedback connections. The final step is to compare the two captions using the Meteor performance assessment metrics. The meteor value is greater in the final caption.

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

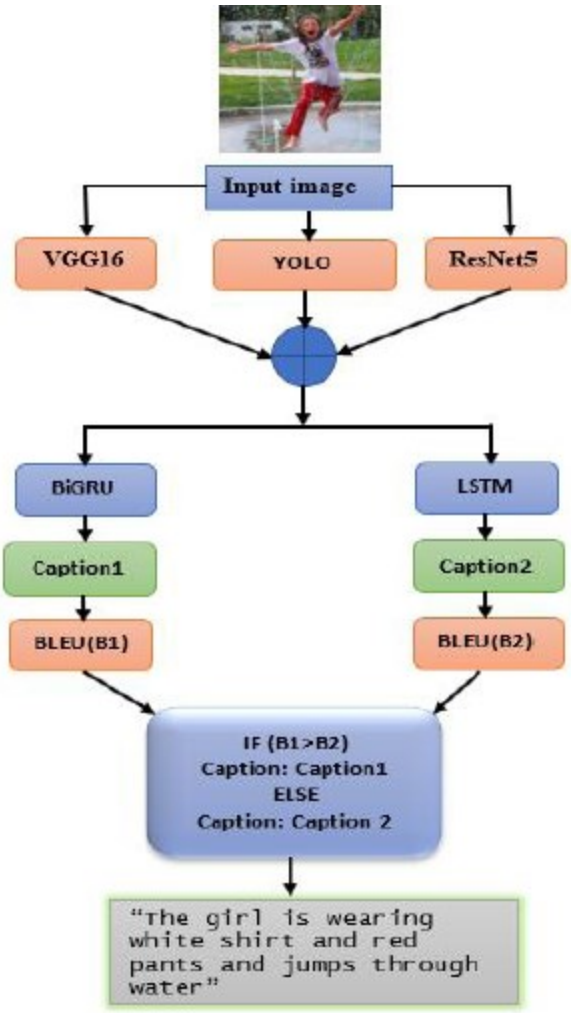

Fig. 4. Proposed image captioning architecture.

IV. DATASETS

All systems that rely on artificial intelligence rely on data. In recent times, image captioning has been fortunate to have access to extensive datasets such as MSCOCO, Flickr8k, Flickr30k, PASCAL, etc., where each picture is accompanied with five associated reference phrases. Various methods and grammars are used to characterize each scene. Microsoft created MSCOCO, a massive dataset whose goal is to characterize the picture as a person. Prior to creating an appropriate caption, it comprehends the scenario and finishes picture identification, segmentation, and generation. It includes 82,783 photos, broken down as follows: 40,504 from the validation set and 40,775 from the test set. There are 28,000 photos used for training, 1,000 for testing, and 1,000 for validation in the Flickr dataset. This research uses Flickr8k as a benchmark dataset for model training. There are eight thousand pictures in all, and each one has five captions that describe the quiet things in detail. All of the photographs include English subtitles that were added by hand. There are two groups within the dataset. There are eight thousand photographs in the image directory, each with five subtitles. We utilize 6,000 of the 8,000 photographs for training, while the other 2,000 are just for fun. The Flickr8k dataset consists of jpg images with resolutions ranging from 256*500 to 500*500. The average phrase length is 12 words.

V. RESULT AND ANALYSIS

Several assessment criteria, including BLEU, METEOR, ROUGE, CIDEr, and SPICE, are used to assess the performance of the picture captions. The BLEU score is used to evaluate the suggested model and compare the predicted words with their original captions. The loss reduced significantly as the number of training epochs

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

expanded, as seen in Fig. 4. It trained our datasets over a longer period of time, 100 epochs to be exact, allowing us to draw more accurate conclusions from our comparisons. Half an epoch to one-hundredth of an epoch is the loss value. Results show a maximum of 10 epochs for losses of 0.5+ and a minimum of less than 0.1 epoch. Figure 5 shows a visual depiction of the BLEU score comparing the predicted caption to five other original captions. The graph shows little fluctuations until the 50th epoch, after which there is a dramatic rise from 0.50 to 0.56 BLEU score from 5 to 10 epochs. "Match words" is an additional metric that measures how many words correspond to the generated text of an image. The graphic depicts the match words going through a dramatic upswell of modifications over time. As shown by 0.49 matching words for 50 epochs and 0.40 for 5 epochs. Both the Match Word and the BLEU Score dipped before they reached their peaks, according to the comparison. For Match words, the score went increased from 0.500 to 0.555 between the fifth and tenth epochs. After then, this sample went through 50 epochs with just little variations, eventually reaching 0.575. The BLEU score showed two separate peaks at the 15 and 30 epochs, with values of 0.450 and 0.470, respectively. A little dip (0.460) appeared in the graph at 35, and at 50, it reached the score (0.480).

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

Fig. 5. Image captions generated by proposed approach.

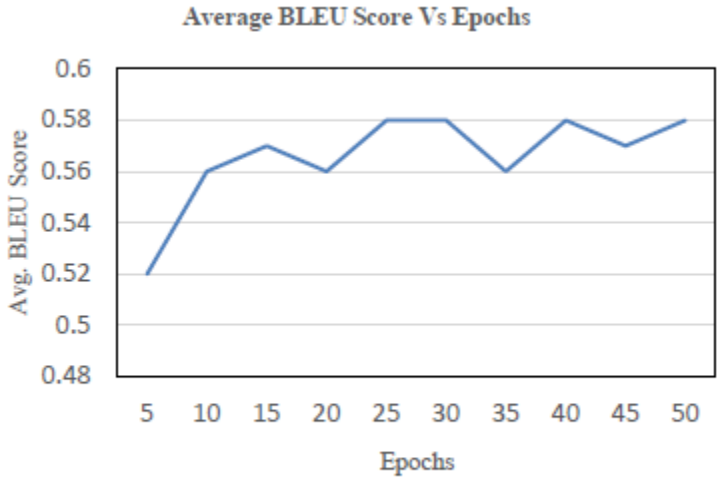

Fig. 6. Average BLEU Score vs Epochs.

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

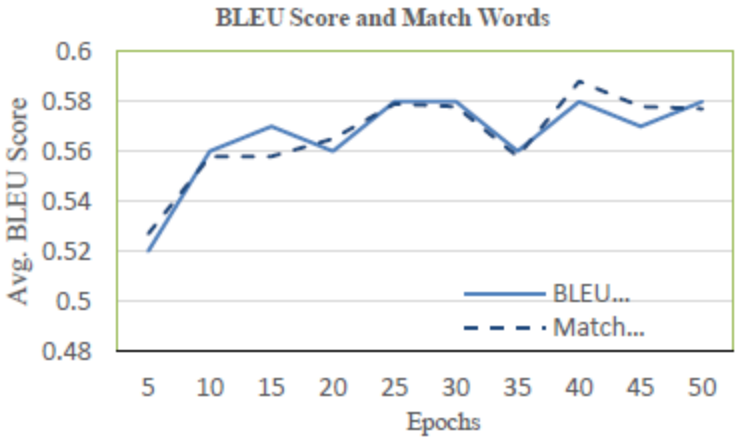

Fig. 7. BLEU Score vs Match words.

The model's recall varies with different threshold levels, as seen in the figure. The threshold values stayed at 1 across the range of 0.0 to 0.25. Following this, there was a gradual decline from 0.25 to 0.75, which got close to the 0.0 value; nevertheless, there was also a little gain of around 0.1 recall value, and the final remembered number was 64.056. A steep peak remains constant at 0.500 accuracy up to the 0.0 to 0.25v threshold value, after which there is a straight increase to 0.675 accuracy, and then a similar value decline up to the 0.75 threshold value (Fig. 6). This graph shows the variance in accuracy with threshold values. The final score for precision is 67.052. Variations in threshold settings and model accuracy levels are shown on the graph. Despite a total precision value of 68.138, a threshold value of 0.75 is achieved by simply increasing the precision values, as opposed to the previous 0.2. In addition to 0.0–0.25, other possible beginning and ending numbers were 1.0–0.25 and 0.5–0.0–0.25. The suitable score is further shown in Fig. 7 by comparing the BLEU score with the Match score. On 5 epochs, the first average score for both is.52. If you wait 10 epochs, the values will rise to 0.56. Overfitting causes it to perform poorly after 35 epochs, after which it reaches its peak performance after 30 epochs. Accuracy and precise recall are shown in Figures 8, 9, and 10.

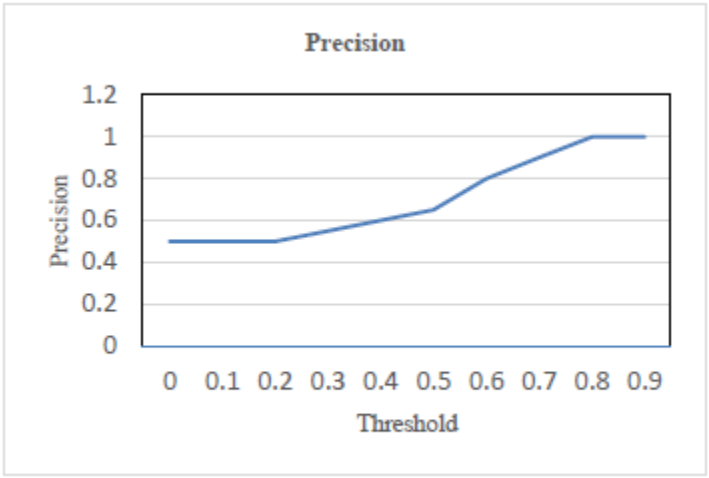

Fig. 8. Precision of proposed systems.

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

Fig. 9. Recall of proposed systems.

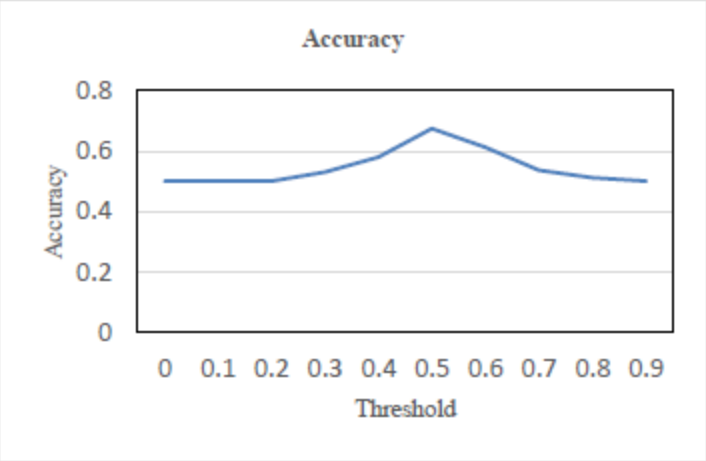

Fig. 10. Accuracy of proposed systems.

The loss and epochs are shown by the shown graph. The given scale indicates that on 0.0 epochs, maximum values are achieved by 1.0. At the end of the first period, the loss was 0.75. Continuing with a graph that shows the loss as a function of time, we see that the loss came to a halt at 17.5 epochs, when the loss value was 0.3.

TABLE I COMPARATIVE ANALYSIS OF PROPOSED APPROACH WITH SINGLE MODEL

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR
Inception V3	0.65	0.43	0.29	0.17	0.21	0.41
VGG16	0.66	0.38	0.30	0.16	0.23	0.22
Res Net50	0.56	0.31	0.18	0.12	0.27	0.51
VGG19	0.61	0.35	0.28	0.18	0.21	0.22
Proposed Hybrid Approach	0.67	0.46	0.35	0.26	0.31	0.54

TABLE II. COMPARATIVE ANALYSIS OF PROPOSED APPROACH WITH HYBRID MODEL

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR
Densenet169 + LSTM [34]	63.73	45.00	30.87	21.13	46.41	19.95
Resnet101 + LSTM [35]	62.77	44.11	30.62	21.10	43.54	18.79
VGG-16 + LSTM [36]	60.56	41.98	28.66	19.51	44.82	19.04
Densenet121 + Attention + LSTM[34]	65.00	46.99	32.83	22.56	47.57	20.44
ResNet152 + Attention + LSTM [37]	65.26	47.55	33.72	23.67	47.54	20.94
VGG-16 + Attention + LSTM [36]	63.81	45.77	32.35	22.55	46.72	20.19
Proposed Hybrid Approach	0.67	0.46	0.35	0.26	0.31	0.54

The provided Tables I and II show the outcomes of a signal encoder-based LSTM decoder model applied to the flickr8k dataset. Inception V3, Res Net50, VGG19, and the proposed hybrid approach are the five encoders shown in the table chart. Each encoder represents a different value of BLEU-1, BLEU-2, BLEU-3, BLEU-4, ROUGE-L, and METEOR. According to the BLEU-1 statistics, the suggested Hybrid Approach Encoder has a maximum value of 0.67. Nevertheless, Res Net50 retains the lowest value in BLEU-2. Using the data from BLEU-3 and BLEU-4, the suggested Hybrid Approach Encoder achieves the highest possible score, whereas ResNet50 sends the lowest scores of 0.18 and 0.12, respectively. In ROUGE-L, the data for Inception V3, VGG16, Res Net50, VGG19, and the Proposed Hybrid method are sequentially numbered as 0.21, 0.23, 0.27, 0.21, and 0.31, respectively. Conversely, when it came to METEOR, the value that was comparable to VGG16 and VGG19 was 0.22.

VI. CONCLUSION

This work presents a technique that uses the Flickr8k dataset to create excellent picture captions using a hybrid encoder-decoder architecture. The suggested approach extracted picture features during the encoding phase using a transfer learning-based model, such as VGG16, ResNet5o, and YOLO. To merge the features and get rid of the duplication, a concatenate function is used. The whole picture caption is obtained during decoding using BiGRu and LSTM. Both the BiGRU and LSTM captions are further tested for BLEU value. With a high METEOR score, the final caption is taken into consideration. Both METEOR and ROUGE assess the suggested model as well. On the Flickr8k dataset, the suggested model obtained BLUE-1: 0.67, METEOR: 0.54, and ROUGE: 0.31. In comparison to other state-of-the-art models, the experimental findings demonstrate that BLUE, METEOR, and ROUGE provide superior outcomes. The model also aids in the real-time generation of the captions.

REFERENCES

[1] J. Gu, G. Wang, J. Cai, and T. Chen, “An Empirical Study of Language CNN for Image Captioning,” Proc. IEEE Int. Conf. Comput. Vis., vol. 2017-October, pp. 1231–1240, 2017, doi: 10.1109/ICCV.2017.138.

[2] J. Aneja, A. Deshpande, and A. G. Schwing, “Convolutional Image Captioning,” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 5561–5570, 2018, doi: 10.1109/CVPR.2018.00583.

[3] K. Xu et al., “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention.” Available: http://proceedings.mlr.press/v37/xuc15.

[4] K. Xu, H. Wang, and P. Tang, “Image Captioning With Deep Lstm Based On Sequential Residual Department of Computer Science and Technology , Tongji University , Shanghai , P . R . China Key Laboratory of Embedded System and Service Computing , Ministry of Education ,” no. July, pp. 361–366, 2017.

[5] S. Liu, L. Bai, Y. Hu, and H. Wang, “Image Captioning Based on Deep Neural Networks,” MATEC Web Conf., vol. 232, pp. 1–7, 2018, doi: 10.1051/matecconf/201823201052.

[6] R. Subash, R. Jebakumar, Y. Kamdar, and N. Bhatt, “Automatic image captioning using convolution neural networks and LSTM,” J. Phys. Conf. Ser., vol. 1362, no. 1, 2019, doi: 10.1088/1742- 6596/1362/1/012096.

[7] C. Wang, H. Yang, and C. Meinel, “Image Captioning with Deep Bidirectional LSTMs and Multi-Task Learning,” ACM Trans. Multimed. Comput. Commun. Appl., vol. 14, no. 2s, 2018, doi: 10.1145/3115432.

[8] M. Han, W. Chen, and A. D. Moges, “Fast image captioning using LSTM,” Cluster Comput., vol. 22, pp. 6143–6155, May 2019, doi: 10.1007/s10586-018-1885-9.

[9] H. Dong, J. Zhang, D. Mcilwraith, and Y. Guo, "I2T2I: Learning Text To Image Synthesis With Textual Data Augmentation."

[10] Y. Xian and Y. Tian, "Self-Guiding Multimodal LSTM - When We Do Not Have a Perfect Training Dataset for Image Captioning," IEEE Trans. Image Process., vol. 28, no. 11, pp. 5241–5252, 2019, doi: 10.1109/TIP.2019.2917229.

[11] K. Xu, H. Wang, and P. Tang, "Image Captioning With Deep Lstm Based On Sequential Residual" Department of Computer Science and Technology , Tongji University , Shanghai , P . R China Key Laboratory of Embedded System and Service Computing , Ministry of Education ," no. July, pp. 361–366, 2017.

[12] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, "Explain Images with Multimodal Recurrent Neural Networks," pp. 1–9, 2014, [Online]. Available: http://arxiv.org/abs/1410.1090.

[13] W. Cui et al., "Landslide image captioning method based on semantic gate and bi-temporal LSTM", ISPRS Int. J. Geo-Information, vol. 9, no. 4, 2020, doi: 10.3390/ijgi9040194.

[14] H. Dong, J. Zhang, D. Mcilwraith, and Y. Guo, "I2T2I: Learning Text To Image Synthesis With Textual Data Augmentation."

[15] C. Liu, F. Sun, and C. Wang, "MMT: A multimodal translator for image captioning," , Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 10614 LNCS, p. 784, 2017.

[16] Q. You, H. Jin, Z. Wang, C. F.-P. of the I., and undefined 2016, "Image captioning with semantic attention," openaccess.thecvf.com Available: http://openaccess.thecvf.com/.

[17] X. Liu, Q. Xu, and N.Wang, "A survey on deep neural network-based image captioning" ,The Visual Computer, 35(3):445– 470, 2019.

[18] A. Farhadi, M. Hejrati, M. Amin Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, "Every picture tells a story: Generating sentences from images", In European conference on computer vision, pages 15–29. Springer, 2010.

[19] G. Kulkarni, V. Premraj, V. Ordonez, S, Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg, "Baby talk: Understanding and generating simple image descriptions", IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(12):2891–2903, 2013.

[20] Y. Yang, C. Lik Teo, H. Daum´e, and Y. Aloimonos, "Corpus-guided sentence generation of natural images", EMNLP 2011 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, (May 2014):444–454, 2011.

[21] M. Mitchell, X. Han, J. Dodge, A. Mensch, A. Goyal, A. Berg, K. Yamaguchi, T. Berg, K. Stratos, and H. Daum´e, " Midge: Generating image descriptions from computer vision detections", EACL 2012 - 13th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings, pages 747– 756, 2012.

[22] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg, "Im2Text: Describing images using 1 million captioned photographs", Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011, NIPS 2011, pages 1–9, 2011.

[23] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using amazon's mechanical turk", In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pages 139–147, 2010.

[24] N. Gupta and A. Singh Jalal, "Integration of textual cues for fine-grained image captioning using deep cnn and lstm", Neural Computing and Applications, 32(24):17899– 17908, 2020.

[25] C. Sun, C. Gan, and R. Nevatia, "Automatic concept discovery from parallel text and visual corpora", Proceedings of the IEEE International Conference on Computer Vision, 2015 Inter:2596–2604, 2015.

[26] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics", IJCAI International Joint Conference on Artificial Intelligence, 2015- Janua(Ijcai):4188–4192, 2015.

[27] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan., "Show and tell: A neural image caption generator", Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 07-12-June:3156–3164, 2015.

[28] C. Amritkar and V. Jabade, "Image Caption Generation Using Deep Learning Technique", Proceedings - 2018 4th International Conference on Computing, Communication Control and Automation, ICCUBEA 2018, pages 1–4, 2018.

[29] A. Ghosh, D. Dutta, and T. Moitra, "A Neural Network Framework to Generate Caption from Images", Springer Nature Singapore Pte Ltd., pages 171–180, 2020.

[30] J. Donahue, L. Anne Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description", IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(4):677–691, 2017.

[31] G. Barlas, C. Veinidis, and A. Arampatzis, "What we see in a photograph: content selection for image captioning", The Visual Computer, 37(6):1309–1326, 2021.

[32] R. Mason and E. Charniak, "Nonparametric Method for Data-driven Image Captioning", pages 592–598, 2014.

[33] X. Dong, C. Long, W. Xu, and C. Xiao, "Dual graph convolutional networks with transformer and curriculum learning for image captioning", arXiv preprint arXiv:2108.02366, 2021.

[34] H., Gao, Z. Liu, L. Van Der Maaten, and Kilian Q. Weinberger, "Densely connected convolutional networks", In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700-4708. 2017.

[35] He, Kaiming, X. Zhang, S. Ren, and J. Sun," Deep residual learning for image recognition", In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. 2016.

[36] Simonyan, Karen and Zisserman, Andrew, "Very deep convolutional networks for large-scale image recognition", arXiv preprint arXiv:1409.1556, 2014.

[37] He, Kaiming, X. Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition", In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. 2016.