Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 46 (2025), No. 1, pp. 81-89
DOI : 10.47974/JIOS-1854

Analysis of attention based deep learning approach for audio image descriptions

Achin Jain [@]
Sarita Yadav [†]
Neetu Singh [§]
Kajal Kaul [‡]
Arun Kumar Dubey *
Prakhar Priyadarshi [#]
Prabhav Sanga [$]
Surinder Kaur [^]
Department of Information Technology
Bharati Vidyapeeth's College of Engineering
Paschim Vihar 110063
New Delhi
India

Ashima Airan [&]
Department of Electrical and Electronics Engineering
Bharati Vidyapeeth's College of Engineering
Paschim Vihar 110063
New Delhi
India

Abstract

The aim of this paper is to focus on utilizing deep learning approaches for the building of an intelligent image captioning system. To bridge the semantic gap between the literal depiction of the images and the languages, this combined architecture is employed. The

* *ORCID-ID:* 0000-0002-6844-9213

[@] *E-mail:* achin.mails@gmail.com
[†] *E-mail:* sarita1320@yahoo.co.in
[§] *E-mail:* singhneetu4bvcoe@gmail.com
[‡] *E-mail:* kajalkaulphd@gmail.com
* *E-mail:* arudubey@gmail.com (Corresponding Author)
[#] *E-mail:* prakharpriya@gmail.com
[$] *E-mail:* prabhav14.sanga@gmail.com
[^] *E-mail:* thisissurinderkaur1304@gmail.com
[&] *E-mail:* ashimajain046@gmail.com

structure of the model is of an encoder-decoder architecture which consists of InceptionV3 CNN for feature extraction and LSTMs with attention mechanism to generate the captions. The complexity of the model is determined based on the usage of the Flickr8K image dataset and it was revealed through the results that it can generate accurate, interesting and appropriate content describing the images. This study has good prospects for the development of framework for learning with multiple modalities especially for usage databases that require content-based searching, image-based data retrieval and usability for users with sight problems. The results show that InceptionV3 Model performs best with BLEU Score of 0.8504 followed by DenseNet 201 with score of 0.8274. Further analysis shows that the best BLEU score of 0.9172 is attained when the dataset of the best performing model, Inception V3, is split into 75:25 Train:Test and running the model for 50 epochs is done. This resulted in a loss function value of 0.70 for the above-mentioned configuration.

Subject Classification: *68T07.*

Keywords: *Image captioning, Inception V3, Deept learning, Audio description.*

1. Introduction

The enormous amount of visual content that users are currently uploading on social networking sites has become a problem to the consumers who tend to comprehend visuals as depicted. This issue is most pressing for the cases of those who are dependent on audio information. They have visual limits which call for more efficient and fast means which will enable insightful images to be captioned in the shortest time possible. This problem is solved for instance by image captioning, which is a relationship between natural language processing and computer vision [1-3]. It enhances the understanding of images and the interaction of a wide range of users towards images by providing detailed and appropriate context anchorage on the images. The adoption of a novel architecture comprising of the CNN-RNN image captioning model constitutes a relevant benefit in the realm of artificial intelligence[4-6]. The goal of this paper is to change the way people understand and describe pictures by combining neural networks with recurrent and attention mechanisms for accurate captions and convolutional networks for image features. This will allow us to create more accurate and intuitive human-like image captions. Picture captioning may be applied to much more than just creating descriptions, it can be used for anything from helping picture search engines to content-based image retrieval systems. The Structure of the paper is as follows: First section is an introduction. The second section reviews earlier research on Image Captioning and provides an overview of the methods we employed. The research technique is then presented in

section three, which also includes a brief overview of the dataset and the suggested approach. The results and discussion of the suggested approach are covered in section four. Finally, section five discusses the conclusion.

2. Related Work

Image captioning studies that were recently conducted have had a major growth in the field. Current models using the importance of the correct areas of the input that got most of their power from the input which in this case is from the input image, which proved that the attention processes, such as focusing on the relevant parts of the input, can be most effectively used for ensuring that the resulting captions are correct and coherent. In their work, Karpathy and Fei-Fei [10] put stress on the fundamentality of visual representations when writing the descriptions. The linking of the visual and the verbal parts is the primary job in image captioning. The inter-modal model is the mechanism to make the model capture the combination of the image with the right captions.Krejtz, Izabela, et al. [8] has shown use of attention description for children.It is new thing for visual impaired person.For instance, Al-Malla et al. [9] introduced an attention-based Encoder-Decoder image captioning model that uses two methods for feature extraction where an image CNN and object features from the YOLO model are employed. The model demonstrated that the combination of object and visual information can most effectively be done. Besides, Chaoyang Wang et al. [4] carried out a review of the literature on picture captioning the respective materials inspecting the task of image captioning and application scenarios and the image captioning algorithms which were based on a template and had an end-to-end encoder-decoder architecture. Al-Shamayleh et al. [7] underlined the vital role of deep learning and attention mechanisms in picture captioning, and giving a detailed survey of the subject's status was its main achievement.

Bahdanau et al. [6] reports that attention mechanisms are the agencies that drive the RNN models to zoom in on vision domains throughout the captioning process, hence, this improves the models' ability to link language elements and visual cues. The best conceivable example of this situation is when Kalchbrenner et al.[5] enhanced their model by making the model to be aware of the visual sequence of pictures. More clearly this means that the model focuses on the specific part in the picture that is relevant to the word being produced at any moment in the caption. The attention-based method has been a tool for generating captions from

images. A deep visual-semantic alignment approach presented by Karpathy and Fei-Fei,[10] encourages more efficient caption creation through parallel learning, which captures the representations of the image as well as the associated descriptions into a common multimodal space. The model had a BLEU-4 score of 27.7, which was considered superior at that time, on the COCO list. A work of Luo et al [11], which encapsulate the trends of the domain, have discussed the fed research efforts in language translation.

3. Research Methodology

3.1 *Dataset*

In this paper, the Flickr8K dataset [12] a collection of 8,000 photos as shown in Figure 1. from the photo-sharing website Flickr is used. Each image has many human annotations on it. There are five different captions for every image in the collection, for a total of 40,000 captions. The pictures show a collection of different scenarios, objects, and activities that occur in everyday life, in cities as well as natural settings.

3.2 *Experimental Model*

Captioning images with multiple layers is the technique used in this study, which starts with text preparation to ensure that we get the best possible input for model training. Before preprocessing, this stage includes text cleaning that is a critical step, building models or plans to guarantee the best input for the caption creation process by eliminating noise and unnecessary information from textual data. The clean text is then separated into individual tokens or words via tokenization, which facilitates the transformation of textual data into a format that machine learning algorithms can process. The approach is shifted to feature extraction from images right after preprocessing of the textual input which is a key feature in image captioning tasks. To achieve this, a Convolutional Neural Network (CNN) model that has already been trained will be the main tool used, seeking the network's capacity to do the job by itself, by extracting crucial and hierarchical visual features from the input photos. Three essential elements are the model architecture constructed for this study, and when put together, they offer a detailed and thoughtful explanation of images.

- **Encoder:** The InceptionV3 convolutional neural network is the model's encoder module. It stands as an ultra-state association (CNN) that is famous for its remarkable job of classifying pictures.

Figure 1

Sample Snapshot of Flick8K Dataset

The encoder can collect a multitude of visual features from the beginning photos.

- **Decoder**: On the other hand, the LSTM network, which is a kind of recurrent neural network (RNN) famous for the ability of regulating sequences, is through the model's decoder module. The caption is the output of the LSTM-based decoder.
- **Attention Model:** One of the important building blocks of the model's architecture is the attention mechanism which allows the decoder to dynamically focus on different parts of the input image one after the other, while producing each word of the caption. The attention model enables the decoder to choose to attend to those areas of the picture that are most significant for producing the next word.

The proposed image captioning model utilizes a CNN-based encoder to produce a static feature vector representing the input image, which is then passed to an attention-based decoder along with the hidden state. The decoder preprocesses the input sequence, embeds it, and concatenates it with the attention-based context vector before feeding it to a GRU. The GRU generates an output and hidden state, which are used to predict the

most probable word from the vocabulary through a dense layer. Figure 2 shows the complete flowchart of the model.

4. Results and Discussion

In this paper, we have carried out comparative analysis of 5 Deep Learning Models: ResNet-50, Inception V3, MobileNet, DenseNet-201, and EfficientNet B3. Table 1 and Figure 3 show the Bilingual Evaluation Understudy score (**BLEU**) score comparison. From the table we have concluded that InceptionV3 outperforms other models with a score of **0.8504**. This result emphasizes how well the Inception V3 model can extract and analyze visual and semantic information from pictures, allowing it to provide contextually relevant and linguistically appropriate captions.

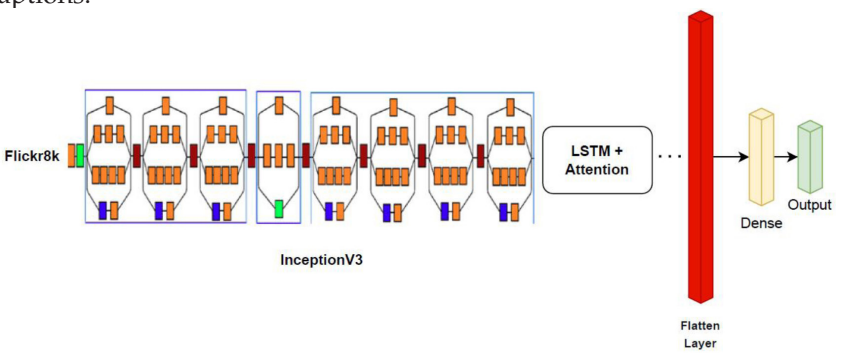

Figure 2

Fine-tuned InceptionV3 model

Table 1

BLEU Score Comparison for Different Models

TL Models	BLEU Score
MobileNet	0.7146
ResNet50	0.7763
InceptionV3	0.8504
DenseNet 201	0.8274
EfficientNet B3	0.7483

Further we have carried out in-depth analysis of InceptionV3 model by running the model with different Train:Test Dataset split ratio for

different epochs. From Table 2 and Figure 4(In Log Scale) it is clear that loss function value for different scenarios and 75:25 Train:Test Split with 50 epochs results in minimum loss function value of **0.70**. This suggests that longer training times can result in higher accuracy and convergence for the Inception V3 model in image captioning tasks. Figure 5 shows the loss plot for 75:25 Split ratio for InceptionV3 Model.

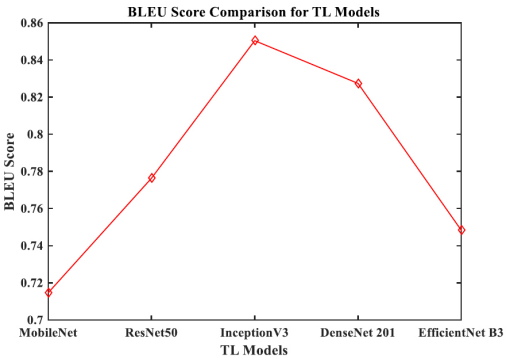

Figure 3

BLEU Score Comparison for TL Models

Table 2

Loss function values for different train-test ratios and for different number of epochs

Epoch / Train: Test Splt	5	15	25	50
10:90	7.94	4.23	3.14	2.58
25:75	6.99	5.06	4.11	2.64
50:50	6.12	5.29	4.48	3.20
75:25	2.20	1.35	0.99	**0.70**
80:20	2.96	1.80	1.32	0.92
90:10	6.68	4.04	2.98	2.11
99:1	77.96	46.31	53.29	57.46

5. Conclusion

This study's comparative assessment of deep learning models for picture captioning has provided important new information about the

functionality and efficiency of the systems. Its greater capacity to acquire and analyze the semantic and visual characteristics of images was demonstrated by the Inception V3 model. Additional refinement of the Inception V3 model using different training epochs showed that longer training times—up to 50 epochs—can result in higher accuracy and convergence. It was shown by examining various train-test data split ratios that a 75-25 split ratio was an efficient way to balance model evaluation and training, which improved performance and generalization. The picture captioning model's high BLEU scores—an average of 85.04—showcase its ability to generate linguistically correct and contextually relevant captions. Furthermore, the inclusion of audio output features in the picture captioning model improves user experience and accessibility, especially for those who are blind or visually impaired. Overall, the results of this work emphasize the remarkable performance of the Inception V3 model and the significance of model optimization through meticulous experimentation with training parameters and data split ratios, contributing to the continuous improvements in the field of picture captioning.

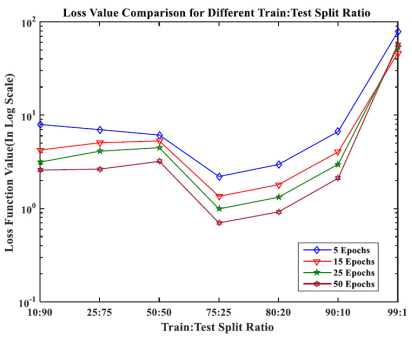

Figure 4

Loss Value Comparison for Different Train:Test Split Ratio

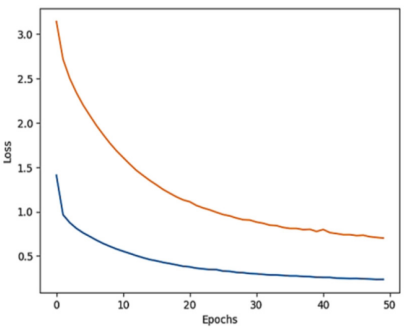

Figure 5

Loss plot for 75:25 Split ratio for InceptionV3 Model

References

[1] A. Vaswani, S. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," Advances in Neural Information Processing Systems 30 (NeurIPS), Long Beach, USA, pp. 5998-6008 (2017).

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," (2018), arXiv preprint arXiv:1810.04805.

[3] I. Sutskever, O. Vinyals, and Q. Le, "Sequence to sequence learning with neural networks," (2014), arXiv preprint arXiv:1409.3215.

[4] K. Cho, B. van Merriënboer, D. Bahdanau, F. O. P. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," (2014), arXiv preprint arXiv:1406.1078.

[5] N. Kalchbrenner and P. Blunsom, "Recurrent continuous translation models," Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (2013).

[6] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," (2014), arXiv preprint arXiv:1409.0473.

[7] A. S. Al-Shamayleh, O. Adwan, M. A. Alsharaiah, A. H. Hussein, Q. M. Kharma, and C. I. Eke, "A comprehensive literature review on image captioning methods and metrics based on deep learning technique," *Multimedia Tools and Applications*, vol. 83, no. 12, pp. 34219-34268 (2024).

[8] I. Krejtz, A. Krejtz, W. Sienkiewicz, and M. Zubek, "Audio description as an aural guide of children's visual attention: evidence from an eye-tracking study," Proceedings of the Symposium on Eye Tracking Research and Applications (2012).

[9] M. A. Al-Malla, A. Jafar, and N. Ghneim, "Image captioning model using attention and object features to mimic human image understanding," *Journal of Big Data*, vol. 9, no. 1, p. 20 (2022).

[10] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3128-3137 (2015).

[11] Y. Luo, J. Li, Z. Li, and Y. Wang, "Natural language to visualization by neural machine translation," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 1, pp. 217-226 (2021).

[12] "Flickr8K dataset," accessed from https://www.kaggle.com/datasets/adityajn105/flickr8k.

Received August, 2024