iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VII **Month of publication:** July 2022

DOI: https://doi.org/10.22214/ijraset.2022.45638

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 10 Issue VII July 2022- Available at www.ijraset.com

Image Captioning - A Deep Learning Approach

Uday Jain[1] , Vansh Goel[2]
[1, 2]*Student, Computer Science and Engineering, Bharati Vidyapeeth College of Engineering New Delhi*

Abstract: ***Image captioning is a brand-new study area in the science of computer vision. The primary goal of picture captioning is to create a natural language description for the input image. In recent years, research on natural language processing and computer vision has become increasingly interested in the problem of automatically synthesising descriptive phrases for photos. Image captioning is a crucial task that demands both the ability to create precise and accurate description phrases as well as a semantic understanding of the images. Long Short Term Memory (LSTM) is used to precisely organise data using the available keywords to form meaningful sentences. The authors of this research propose a hybrid system based on multilayer Convolutional Neural Networks to create a lexicon for characterising the visuals. The convolutional neural network employs trained captions to deliver an accurate description after comparing the target image to a sizable dataset of training images. We demonstrate the effectiveness of our suggested methodology using the Flickr 8K datasets.***

Keywords: Deep learning, LSTM, neural network, glove embedding, image captioning, ResNet-50

I. INTRODUCTION

Image captioning is a deep learning issue in which the goal is to come up with a caption for a given input image. It makes use of a combination of computer vision and natural language processing algorithms.

We employed a deep learning-based model in this research, which consists of a combination of LSTM and CNN (Transfer Learning) networks. Our project's goal is to train the algorithm to correctly predict captions for supplied images. Because of the vast size of the flickr 30k dataset,flickr 8k dataset has been used .

The model's accuracy is restricted by the amount of the dataset we use and the number of variables we employ.

Specifications used : NVidia Tesla P100 GPU with 16GB vRAM on Kaggle, along with 13 GBsystem RAM.

The Flickr 8k dataset [6] has 8092 captioned photographs, each with five different captions. The token.txt file has a total of 40460 captions. Each caption gives a detailed account of the entities and events depicted in the image. There are 6000 photos in the dataset for training, and 1000 images each for testing and validation. The data set includes a number of different types of data. For obtaining vector representations for words, GloVe Embedding [7] is employed. Thevector representation of words is stored in the glove embeddings. We utilised the 6B 100d version, which has 6 billion words and 100 vectors for each of them. It is an open source NLP project developed by Stanford University.

Generated caption - two chairs on the beach.

Figure 2 : A visualization of self-attention in our proposed Object Relation Transformer.

In relation to the chair that is highlighted in red, the transparency of the identified object and its bounding box varies with the attention weight. Our model shows linkages in the generated caption between this chair and the companion chair to the left, the beach below them, and the umbrella above them.

II. RELATED WORK

Since the creation of the Internet and its extensive use as a platform for sharing photographs, the issue of image captioning has persisted, as have the solutions that have been suggested. Many algorithms and methods have been proposed by researchers from various angles. Wang [19] and Q. Wu, C. Shen, and P. Wang [10] employed two LSTM units to construct a skeleton phrase and then add extracted properties to it. Sub-blocks or parts of images are used as features by K. Fu,

J. Jin, R. Cui, F. Sha, and C. Zhang [9], who construct descriptions for those features. Each caption for an image is given a likelihood by Bo Dai and Dahua Lin [11], and the caption with the highest probability is chosen.For producing context-aware captions, R. Vedantam, S. Bengio, and K. Murphy [18] employed the discriminator class. Instead of captions, Z. Gan [14] employed semantic notions found in the image. Chuang [12] applied the same algorithm to three LSTM units - To produce captions, use parameters but different styles (such as facts, humour, and so on. The attributes of different parts of an image were supplied into the LSTM for caption generation by L. Yang and K. Tang [16]. J. Krause, J. Johnson, and R. Krishna [17], like [16], employed picture areas to produce phrases using hierarchical RNN. To create captions, S. Venugopalan and L. A. Hendricks [15] and F. Liu, T. Xiang, and T. M. Hospedales

[18] employed several training datasets and varied techniques in different layers of the network.Cortana is a Microsoft virtual assistant; their most current research focuses on merging AI with Bing's search service, allowing users to interact more organically. Another method for captioning images that attempts to tie the words in the anticipated caption to specific locations in the image is the family of attention-based techniques [26, 30, 28]. The spatial localisation is constrained and frequently not semantically relevant because the visual attention is frequently taken from higher convolutional layers of a CNN. Anderson et al. in [2] addressed this restriction of conventional attention models by fusing a "bottom-up" attention model with a "top-down" LSTM, which is most comparable to our approach. The Faster R-CNN object detector's recommended regions of interest are the focus of the bottom-up attention, which is applied to mean-pooled convolutional features [20]. The top-down LSTM is a two-layer LSTM, with the first layer acting as a visual attention model to focus on the pertinent detections for the current token and the second layer acting as a linguistic LSTM to produce the next token. Using this method, the authors achieved cutting-edge performance for visual question answering and image captioning, highlighting the advantages of fusing information generated from object identification with visual attention. Again, the use of spatial information, which we advocate in this work through geometric attention, was not made. Hu et al. introduced geometric attention for object detection for the first time in [9]. There, the authors inferred the significance of the association between pairs of objects using bounding box coordinates and sizes. They made the premise that if two bounding boxes are nearer to one another and more comparable in size, then their relationship is stronger.

III. DATASET AND MODEL OF DESIGN

Several annotated picture datasets are available for the task of captioning photos. The most popular ones are MSCOCO Dataset, Flickr 8K, and Pascal VOC Dataset. In the suggested model, the Flickr 8K Image Captioning dataset [9] is utilised. A dataset known as Flickr 8K consists of 8,092 pictures taken from the Flickr.com website. This dataset includes a collection of daily activities and the captions that go with them. Each object in the image is first given a label, and then a description based on those objects is added. From this corpus, we divided the 8,000 photos into three separate groupings. The development and test datasets each contain 1000 photos, whereas the training data (DTrain) has 6000 images.

Multimodal recurrent and convolutional neural networks are the foundation of our model for captioning images. The features of a picture are first extracted using a convolutional neural network, and the captions and features are then fed into a recurrent neural network. Figure 2 depicts the architecture of the picture captioning model.

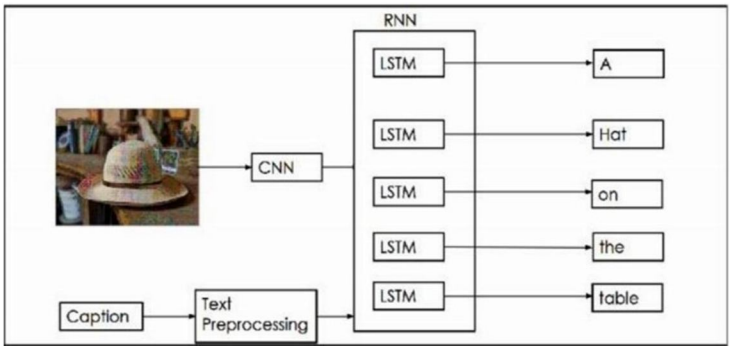


Figure 2 : Architecture

The model has three phases

A. *Extraction of Image Feature*

Due to the model's success in object identification, the VGG 16 model is used to extract the features of the photos from the Flickr 8K dataset. The VGG is a 16-layer convolutional neural network that follows a pattern of 2 convolutional layers, 1 dropout layer, and then a fully connected layer at the end. Since this model configuration learns relatively quickly, the dropout layers are there to prevent the training dataset from becoming overfit. These go through a Dense layer's processing to create a 4096-vector element representation of the image, which is then forwarded to the LSTM layer.

B. *Sequence Processor*

In order to handle the text input, a sequence processor serves as a word embedding layer. The embedded layer includes a mask to disregard padding data and algorithms to extract the necessary text features. The last step in the picture captioning process is connecting the network to an LSTM.

C. *Decoder*

The model's final phase combines the input from the Image extractor and Sequence Processor phases using an additional operation before feeding it to a 256 neuron layer and finally to a final output Dense layer. This produces a softmax prediction of the next word in the caption over the entire vocabulary, which was formed from the text data that was processed in the Sequence Processor phase. Figure 3 depicts the network's organisational structure for comprehending the text and image flow.

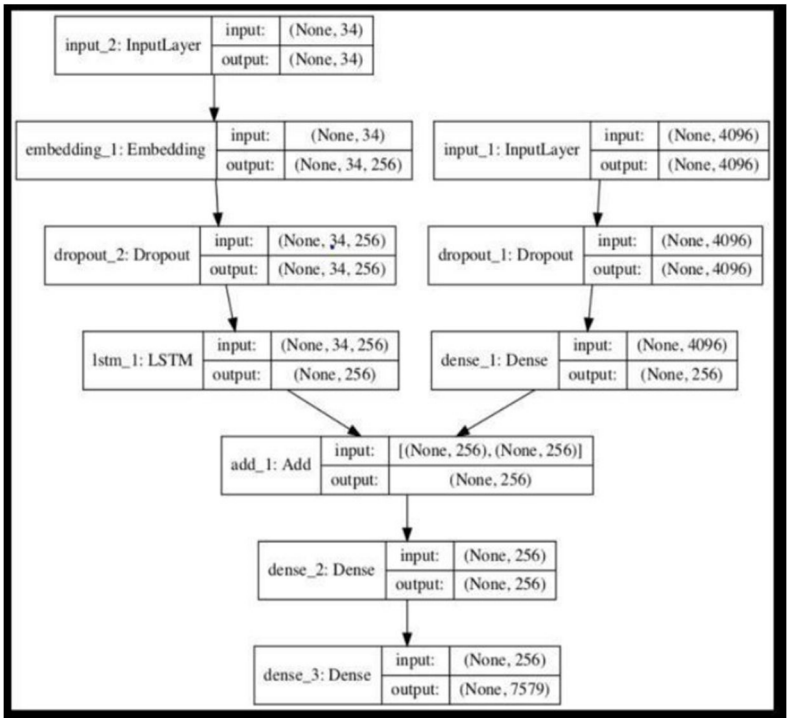


Figure 3 - Image Captioning Model

IV. EXPERIMENTAL RESULTS

We have introduced a single joint model based on ResNet50 and LSTM with software attention for automatic picture captioning. One encoder-decoder architecture was used for creating the suggested model. To compress an image into a small representation that can be represented graphically, we used ResNet50, a convolutional neural network. The decoder for the descriptive sentence was then chosen as a language model LSTM. In the meanwhile, we combined the LSTM with the soft attention model so that learning may be targeted at a specific area of the image to enhance performance. Using stochastic gradient descent, which facilitates training, the entire model is fully trainable. The experimental results show that the suggested model can produce quality image captions automatically.

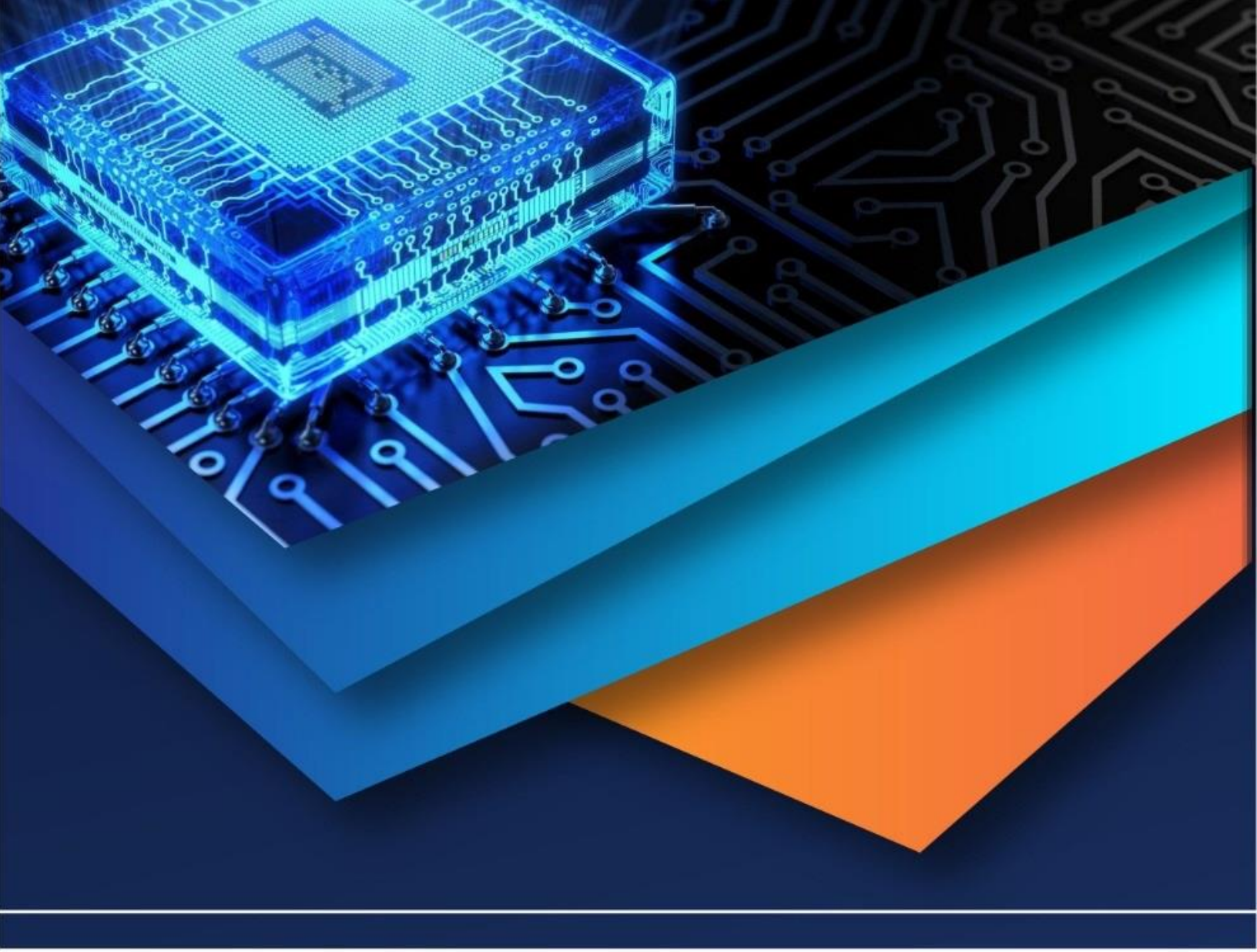
Figure 4 : Results

V. CONCLUSION

The concepts of certain popular picture captioning techniques are introduced and examined in this work. We introduce the necessary data sets and assessment indices for this field. Although the prediction impact of the currently available image captioning algorithms has improved to some level, they do not fully achieve the function of creating specific description statements in accordance with certain conditions. Image captioning can be enhanced in three ways in the future. The first is to make the network more adaptable so that the model may focus on particular situations and issues and produce tailored descriptions in accordance with various circumstances and concerns. The second part involves improving the evaluation algorithm to more properly assess the model's output sequence's quality. The third goal is to make the model more robust and prevent interfering characteristics from having an impact on the model's output.

REFERENCES

[1] Geoffrey E. Hinton, Alex Krizhevsky, and Ilya Sutskever, "ImageNet Classification with Deep Convolutional Neural Networks," http://papers.nips.cc/paper/4824- imagenetclassificationwith-deep-convolutionalneural-networks.pdf [Online]
[2] Li-Jia Li, Kai Li, and Li Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database, Jia Deng, Wei Dong, Richard Socher,
[3] Deep VisualSemantic Alignments for Generating Image Descriptions by Andrej Karpathy, Li Fei-Fei, https://cs.stanford.edu/people/karpathy/cvpr2015.pdf is [online] accessible.
[4] Sadaqat ur Rehman, Yongfeng Huang, Yu-Jin Zhang, Image Captioning with Object Detection and Localization Available online at: ftp:/arxiv.org/papers/1706/1706.02430.pdf Convolutional Image Captioning, Jyoti Aneja, Aditya Deshpande, Alexander Schwing, Available online at: arxiv.org/pdf/1711.09151.pdf
[5] Pinar Duygulu, Jia-Yu Pan, and Hyung-Jeong Yang, Automatic Image Captioning, Multimedia and Expo, 2004, ICME '04, IEEE International Conference on, Volume: 3
[6] Dumitru Erhan, Oriol Vinyals, Samy Bengio, Alexander Toshev, Show and Tell: A Neural Image Caption Generator [Online] Easily accessed at: https://arxiv.org/pdf/1411.4555.pdf
[7] Yoshua Bengio, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, Kyunghyun Cho, Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, Available online at: arxiv.org/pdf/1502.03044.pdf "Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics", Journal of Artificial Intelligence Research, Volume 47, pages 853–89, by M. Hodosh, P. Young, and J. Hockenmaier (2013)
[8] Wang, Z., Fang, C., You, Q., Jin, H., and Luo, J. (2016). semantically accurate image captioning. Pages 4651–4659 of the IEEE conference proceedings on computer vision and pattern recognition.
[9] Luo, J., You, Q., Jin, H., Wang, Z., Fang, et al (2016). semantically accurate image captioning. Pages 4651–4659 of the IEEE conference proceedings on computer vision and pattern recognition.
[10] J. Lu, C. Xiong, D. Parikh, & R. Socher (2017). Understanding when to look: For image captioning, adaptive attention is provided by a visual sentinel. Pages 375–383 of the IEEE conference proceedings on computer vision and pattern recognition.
[11] Liu, W., Nie, L., Shao, J., Xiao, J., Chen, L., Zhang, H., & Chua, T. S. (2017). Convolutional networks for image captioning using spatial and channel-wise attention. Pages 5659–5667 of the IEEE conference proceedings on computer vision and pattern recognition.
[12] Yao, T., Pan, Y., Li, Y., Qiu, Z., and Mei, T. (2017). enhancing the captioning of images with attributes. Pages 4894–4902 of: Proceedings of the IEEE International Conference on Computer Vision [14] The authors are Cornia, Baraldi, and Cucchiara (2019). A methodology for creating controllable and grounded captions is called "show, control, and tell." Pages 8307-8316 of the book Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
[13] Lashin, V., and Rahtu, E. (2020). Dense video captioning with multiple media. Pages 958–959 of the book Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.
[14] Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, et al (2017). Using rich crowdsourced image annotations, the Visual Genome project connects language and vision. 123(1) of the International Journal of Computer Vision, 32–73.
[15] 18. Convolutional Image Captioning, Jyoti Aneja, Aditya Deshpande, and Alexander Schwing Available online at: arxiv.org/pdf/1711.09151.pdf
[16] Yoshua Bengio, Ruslan Salakhutdinov, Aaron Courville, Kyunghyun Cho, Jimmy Lei Ba, Ryan Kiros, Richard S. Zemel, Kelvin Xu, Show, Attend, and Tell: Visual Attention and Neural Image Captioning Available online at: arxiv.org/pdf/1502.03044.pdf [9] "Framing Image Description as a Ranking Task: Data, Models, and Evaluation Metrics", Journal of Artificial Intelligence Research, Volume 47, pages 853-899, by M. Hodosh, P. Young, and J. Hockenmaier (2013)
[17] Samy Bengio, Dumitru Erhan, Oriol Vinyals, Alexander Toshev, Show and Tell: A Neural Image Caption Generator, Available online at: arxiv.org/pdf/1411.4555.pdf Convolutional Image Captioning, Jyoti Aneja, Aditya Deshpande, Alexander Schwing, Available online at: arxiv.org/pdf/1711.09151.pdf
[18] Sadaqat ur Rehman, Yongfeng Huang, Yu-Jin Zhang, Image Captioning using Object Detection and Localization, Available online at: ftp:/arxiv.org/papers/1706/1706.02430.p df
[19] Deep VisualSemantic Alignments for Generating Image Descriptions by Andrej Karpathy and Li Fei-Fei, https://cs.stanford.edu/people/karpathy/cvpr2015.pdf is [online] accessible.
[20] Deep Convolutional Neural Networks for ImageNet Classification by Alex Krizhevsky, Ilya Sutskever, and Georey E. Hinton https://papers.nips.cc/paper/4824-imagenetclassificationwith-deep-convolutionalneura l-networks.pdf [Online]



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY