International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Automated Image Caption Generating System with Caption to Speech Convertor Using Deep Learning Methods

Shubham Rawale[1], Megha Ghotkar[2], Krishna Sonavane[3], Paras Surve[4], S.R.Khonde[5], D.D.Ahir[6]

[1,2,3,4] *Student (Final Year), Department of Computer Engineering, Modern Education Society's College of Engineering, Pune-01, Maharashtra, India*
[5,6] *Professor, Department of Computer Engineering, Modern Education Society's College of Engineering, Pune-01, Maharashtra, India*

ABSTRACT

The field of image processing is vast, and you can always make new discoveries with more accurate results than ever before. This field of research is huge, providing a broader space for future research, and will eventually help improve computers, make them function similarly, and access the world visually. Image captioning technology aims to automatically generate favourite captions by providing image descriptions, so as to use words that provide the meaning of the image to better understand the image. These words provide the image processed by the caption method. The image captioning process needs to accurately identify the objects and scenes provided by the image, and you must have the ability to process and analyse them. Our model successfully performed the process of converting images into vector form and finally predicted subtitles by using the additional function of the g TTS (Text to Speech) engine to convert subtitle text to speech. The system first acquires images and uses CNN to convert them into vectors, and then combines them with word vectors generated by LSTM from a recurrent neural network to finally predict the most realistic image title. This model uses our proposed fusion method to reduce the state of the RNN hidden vector by four times, which helps the system obtain more accurate results. After the Results, the BLEU score was used for performance evaluation.

Keywords: Image Caption,Text to Speech, Deep Learning, Image Processing, CNN, RNN

1. INTRODUCTION

Human beings can easily understand the content of images and describe their environment, and express them in the form of sentences in natural language as needed. But it is very difficult for the machine to describe the image in detail. For the machine, it requires a comprehensive application of image processing, natural language processing, computer vision and other important fields. This field allows humans to obtain information that helps them understand and perceive the vast landscapes around the world. Today, they receive increasing attention in the fields of image processing, computer vision and interaction. This recent development will fuel the next researchers and the latest discoveries. The development of these systems will surely help people to see and understand the world more visually in the future. Our system was trained on the flicker8k data set. The proposed model combines (merges) the image vector produced using CNN and automated feature extraction and partial captions vectors using RNN and LSTM with back propagation algorithm and produces a whole matrix for feeding it forward to the SoftMax layer, making it ready for caption predictions and the system converts the predicted caption text into the speech using g TTS engine, also which is a very useful case for visually impaired and will help them to understand the surrounding well and will get information from the world.

2. RELATED WORKS

[1] In the paper, they used CNNRNN-based image captioning model and reinforcement learning. They have established their evaluation indicators with their three systems.
[2] Dorsey et al. In 2018, he proposed the blind reading method. You use the Google API to extract the image and convert it to JSON format. After converting the text, it will be converted to synthesized speech.
[3] This is the article; they proposed a model based on the encoder and decoder. They use the CNN structure to get the code from the image.

* *Corresponding author.*
E-mail address: shubhamrawale681@gmail.com

[4] Here the machine learning model uses VGG16 and CNN to generate titles and uses RNN to transmit data The system also fetches images according to the user's search query.
[5] Aker and Gaizauskas created an automatic subtitle model using geotagged images and it has a dependent model mode that provides a variety of scenes. They use ROUGE scores for modelling purposes.
[6] Horag uses the attention mechanism to create image caption models, collect larger image data sets, and evaluate the results of machine learning models.
[7] Yagn, gave a system that uses natural language to automatically generate subtitles, and proposed a multi-modal neural layer to locate the human body and receive sequences.
[8] proposed a system that uses LSTM technology and a CNN layer to generate the following word sequence, and obtain subtitles after combining the predicted words.

3. METHODOLOGY AND IMPLEMENTATION

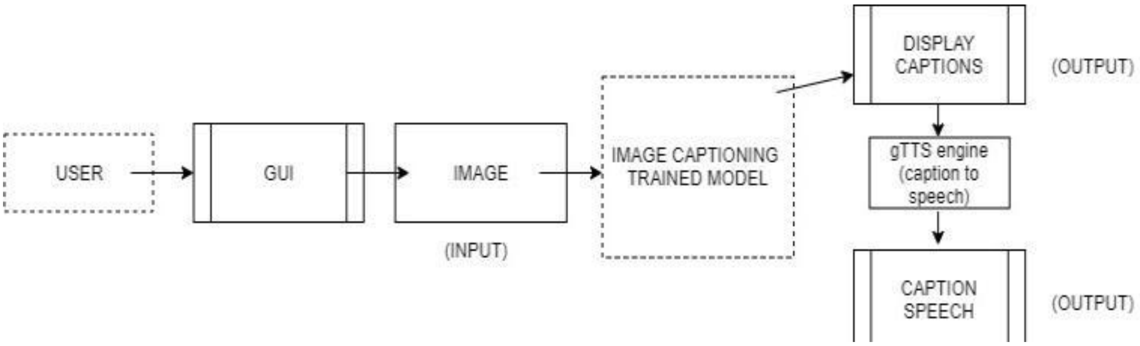

Fig. 1 System Flow Diagram

.
The model was implemented on web platform, where it used python flask for the GUI and its frameworks like Keras for integration purpose by using deep learning methods which produces the captions based on neural networks using recurrent neural network with a long short term memory layers in its hidden state. The Model uses flicker8k dataset which has almost 8000 images with 6000 images being separated for training sets and other 1000 each to development and testing datasets. The images are converted into a fixed size vector before feeding it to the neural network using transfer learning with use of CNN. The image captions of 5 each for particular image are converted into word vector on other hand using recurrent neural network with long short-term memory to enhance the memory time for smooth prediction after training. We framed it as a supervised learning problem wherein we have set of data points as:

Data Points **D = [Xi, Yi]**
(Where Xi is feature vector and Yi is the target variable for framing)

After the processing of a prefix, the RNN passes the target variable for feeding it into SoftMax to predict the next word using forward sequencing. Thus, the data matrix is formed using indexing of the most probable words using sequence processing. Then the batch processing is done for optimization with using back propagation algorithm for cost minimization and the words are selected greedily using greedy search algorithm for forming the sequence. Finally, the SGD gradient which is Adam optimizer is being used to merge the model. After the successful caption prediction by the system, it also then at last converts the generated captions into speech via text to speech conversion engine.

4. MODEL ARCHITECTURE

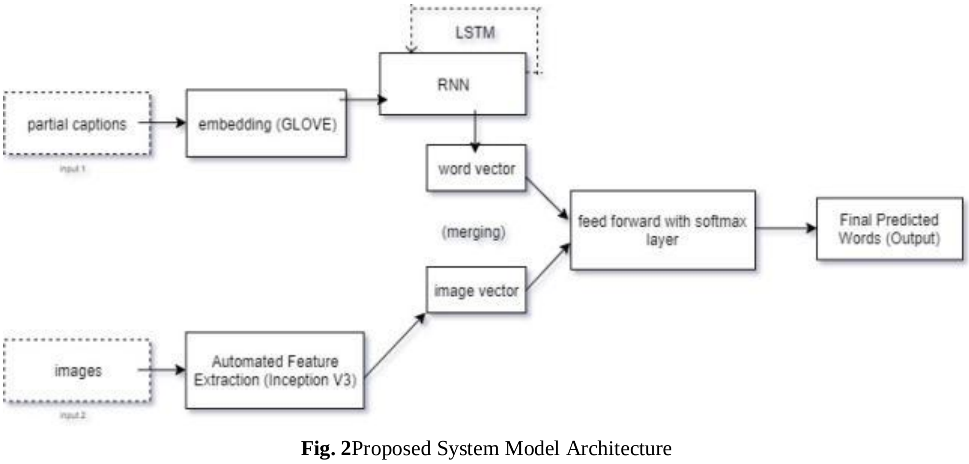
LSTM
RNN
partial captions
input 1
embedding (GLOVE)
word vector
(merging)
feed forward with softmax layer
Final Predicted Words (Output)
image vector
images
input 2
Automated Feature Extraction (Inception V3)

Fig. 2 Proposed System Model Architecture

We proposed a merging model architecture which merges the fixed image vectors and word vector to form a best model which predicts the captions for the given images. As we have two parameters for system input for merging which are feature vectors from the images and the partial captions from the given dataset. The image vector are produced using automatic feature extraction method which uses inception V3 model and the partial vectors are produced using glove model with embedding it to form a word matrix to move towards the feed forwardlayer In the model, RNN does not takes images into its subnet but takes only one word prefix at a moment. The fixed sized vector is then merged together into multimodal layer and available for the SoftMax and finally generates the word sequence using sequential APIs as an output. This words later on are greedily selected which had a maximum probability using Greedy search algorithm. At last, the captions generated as output are given as an input to the g TTS and then it converts these caption texts into the speech format.

5. EVALUATION AND FINAL RESULTS

Fig 3,4

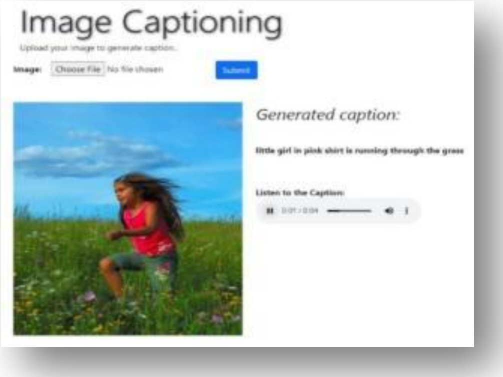

Image Captioning

Upload your image to generate caption..

Image:

Generated caption:

little girl in pink shirt is running through the grass

Listen to the Caption:

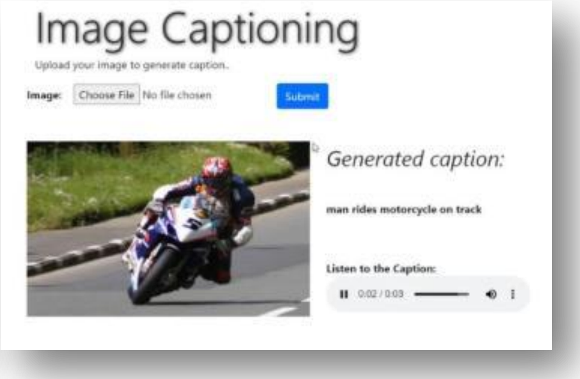

Image Captioning

Upload your image to generate caption..

Image:

Generated caption:

man rides motorcycle on track

Listen to the Caption:

Systems
User Interface and Predicted Captions with Speech Converted Output

We have successfully implemented our system, above are some snippets of the system with its interface and the outputs. We achieved great results from our trained image captioning models with high accuracy than other previous counterparts and used a huge flicker 8k dataset to train our model rigorously and received accurate results using our proposed merging model where the images and their caption vectors are being converted and merged into the SoftMax layer for final sequence generation. Finally, the bleu score was calculated and our system got a score of **0.92425**. with systems accuracy up to **96%** on the pre-testing dataset images and **92%** on random images provided to the system for further predictions.

6. CONCLUSION

Here we conclude our final project implementation on an image captioning system using modern machine learning techniques. The system successfully generates the captions and further converts the captions into speech for better user interaction. Moreover, we have also implemented our new feature of converting the generated image captions into speech which will be a great help for the visually impaired and enables them to understand the images and their surroundings well and at last, the end-user can also download the converted speech as mp3 file for any other uses as well. Our proposed merging model successfully got integrated with the built Python flask GUI which made it a full-fledged application to use. Our system is fully capable of generating captions and converting them into speech.

7. FUTURE SCOPE

The future scope of the field is very broad and is highly dependent on the type of data set used, and very domain-specific. The system produces the best captions that resemble the human mind and its processing. In the future, the field of image processing will easily serve the main purpose of simplifying human life and will become a good interface between humans and their environment to get a better understanding of the world using modern computer technology. The image caption field plays an important role in the study of image processing. It can be widely used in medical, traffic monitoring, computer vision, and other fields, making the system very similar to human thoughts and its thinking as the human brain. The system will have a far-reaching impact in the future, and it can be made more precise and accurate using modern tools and technologies which will appear in the coming future. The system has a great scope in the future and can become more accurate and precise using modern tools and technologies which will come in the future. Finally serving the main purpose of connecting the 3 dots which are human, machine, and their surrounding world.

REFERENCES

[1] Shuang Liu, Image Captioning Based on Deep Neural Networks, MATEC Web of Conferences 232, 01052 (2018) Available: https://doi.org/10.1051/matecconf/201823201052.

[2] Doshi, Text Reader for Visually Impaired Using Google Cloud Vision API, international journal of innovative research in technology (IJIRT). Vol. 4, 5/18.

[3] Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, vol. 39. 4/17.

[4] ShaunakBaradkar, Cap Search - "An Image Caption Generation based search", International Research Journal of Engineering and Technology (IRJET) Vol. 6, 4/19.

[5] An Overview of Image Caption Generation Methods, Hindawi Computational Intelligence and Neuroscience Volume 2020, Article ID 3062706, 13 pages https://doi.org/10.1155/2020/3062706

[6] Zhongliang Yang, Yu-Jin Zhang, Sadaqat ur Rehman, Yongfeng Huang, Image Captioning with Object Detection and Localization, [Online]Available: https://arxiv.org/ftp/arxiv/papers/1706/1706.02430.p df

[7] Image Caption Generator using Big Data and Machine Learning, International Research Journal of Engineering and Technology (IRJET), Vol.7, 4/20.

[8] K. Simonyan and A. Zisserman, ''Very deep convolutional networks for large-scale image recognition,'' 2014.

[9] Karpathy, Andrej, and F. F. Li. "Deep visual-semantic alignments for generating image descriptions." Computer Vision and Pattern Recognition IEEE,

[10] onahue et al., "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, pp. 677-691, 1 April 2017.

[11] Mao, J, Xu, W, Yang, Y, et al.: 'Explain images with multimodal recurrent neural networks', arXiv preprint arXiv, October 2014, 1410.1090

[12] Johnson, J, Karpathy, A, Fei-Fei, L.: 'Fully convolutional localization networks for dense captioning'. Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA, 2016, pp. 4565–4574

[13] Hochreiter, Sepp, and J. Schmidhuber. "Long ShortTermMemory."Neural Computation 9.8: 1735-1780.

[14] Vinyals, O, Toshev, A, Bengio, S, et al.: 'Show and tell: a neural image caption generator'. Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR), Boston, USA, 2015, pp. 3156–3164

[15] S. Yan, F. wu, J. Smith and W. Lu,"Image Captioning via a Hierarchical Attention Mechanism and Policy Gradient Optimization," LATEX CLASS FILES, vol. 14, 11 January 2019

[16] A. Karpathy and L. Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions," CVPR 2015 Paper, December 2014.

[17] K. Simonyan and A. Zisserman, "VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION," in ICLR, 2015.

[18] S. ALBAWI and T. A. MOHAMMED, "Understanding of a Convolutional Neural Network,"in ICET, Antalya, 2017.

[19] G. Ding, M. Chen, S. Zhao, H. Chen, J. Han and Q. Liu, "Neural Image Caption Generation with Weighted Training and Reference," Cognitive Computation, 08 August 2018

[20] Gu, Jiuxiang, et al. "Stack-Captioning: Coarse-to-Fine Learning for Image Captioning." (2018)