Automatic Image Captioning

Pooja Naik[1*], Prajna Bhat[2], G. V. Rakshitha[3], G. S. Roopa[4], K. Athmaranjan[5]

[1,2,3,4]*Student, Dept. of Information Science and Engineering, Srinivas Institute of Technology, Mangalore, India*

[5]*Associate Professor, Dept. of Information Science & Engg., Srinivas Institute of Technology, Mangalore, India*

*Corresponding author: poojapn7@gmail.com

***Abstract*: Image captioning has become a modern and daunting activity that has attracted a widespread attention. The task involves generating a concise description of an image in natural language and is currently accomplished using techniques that combine the processing of natural language with the machine learning methods. In this paper we presented a paper a model which generates a description of an image in natural language. A combination of convolutional neural network was used to extract features and then recurrent neural networks are used to generate text from these features. We used a dataset called Flickr8k. The obtained results are positive and competitive.**

***Keywords*: Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Machine Learning.**

1. Introduction

To human beings a simple look at an image is enough to identify the objects and regions in that image. But it's not the same for computers some previous models simply labelled the objects in the picture. There some models which generates image descriptions but rely on hard coded visual concepts. This imposed constraint on their variety. Text description created automatically for picture is the job of image captioning. Experiments on multiple datasets demonstrate the consistency of the model and the fluency of the of the language it only learns from the explanation of the picture. In this work we strive to take a step towards the goal of producing detailed image descriptions. We introduced multimodal architecture of the Recurrent Neural Network that takes an input image and generates its description. Our experiments show that the sentences produced outperform retrieval based baselines significantly, and produce meaningful qualitative predictions. We then train the model on the inferred correspondences ant assess its performance on a new regions level annotation dataset. The model produces definitions of the image ant its regions in the natural language. This approach leverages image dataset and their description of sentences to learn about the intermodal correspondences between the language and visual data. This alignment model is based on a combination of convolutional neural networks over image regions, bidirectional recurrent neural networks over sentences and a specified goal that aligns the two modalities with a multimodal embedding. It then describes an architecture of the multimodal recurrent neural network that uses the inferred alignments to learn how to generate image region descriptions.

2. Proposed System

CNN features are capable of describing an image. A common approach for applying this ability to natural language is to extract sequential information and turn it into a language. They extract feature map from top layers of CNN in most recent image captioning works, move it on to some sort of RNN and then use a soft max to get word score at every stage. Now our goal is to identify in addition to captioning, the objects in the picture to each word refers. To put it another way, we want details about location. So we need to extract features from a lower level of CNN, encode it into a vector that is dominated by the feature vector referring to the object that the word wants to represent, and move it on to RNN. RNN then generates the natural language description of the image.

3. Literature Survey

Work [1] proposed an architecture of a deep, convolutional neural network codenamed Inception. The key feature of this architecture is the improved usage of the computing resources within the network. But features learned from the lower layers may contain more accurate information about the correlation between captioned words and specific image location.

Work [2] presented a generative model based on a profound recurrent architecture combining advance in computer vision and machine translation that can be used to generate natural phrases describing an image. Given the training image, the model is trained to maximize the likelihood of the target description.

Work [3] implemented a model based on profound, convolutional networks that performed very well in tasks of image interpretation. The recurrent convolutional model and long term RNN models are suitable for end to end trainable.

Large scale visual learning and demonstrate the value of these models on video recognition tasks in benchmark. Attention mechanism has a long history, especially with respect to image recognition. The associated work comprises work [4] and work [5].

Figure 1 shows the interaction between user and system. User will be given an option to insert the image where he can insert any image available in his system. After inserting an image, the system will pre-process the image and the pre-processed image will be fed to trained CNN model. The output of CNN will be given to RNN which generates the caption for the image. This generated caption will be displayed for the user.

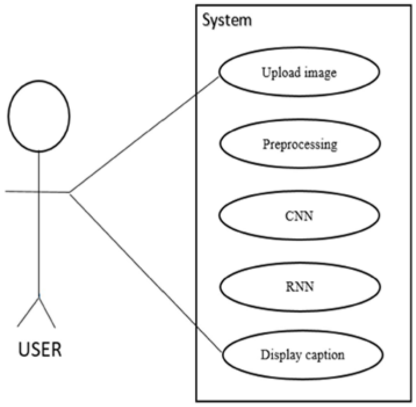

Fig. 1. Use case diagram of automatic image captioning

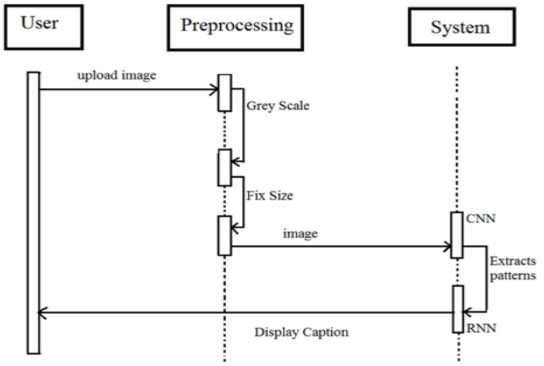

Fig. 2. Sequence diagram of automatic image captioning

Figure 2 shows the sequence of steps. Initially user inputs the image for pre-processing. Here image will be converted to grey scale then it will be converted to fixed size then image is fed to CNN.CNN generated output will be given to RNN. Now RNN generates the caption and that caption is displayed to user.

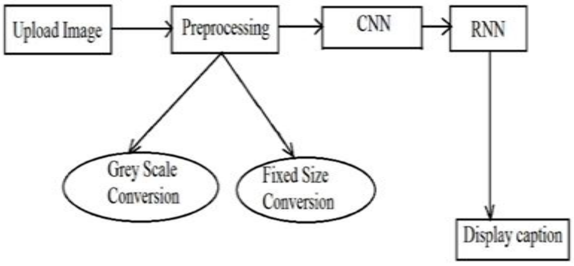

Fig. 3. The data flow diagram for automatic image captioning

Figure 3 shows Data flow diagram. This shows different stages of how an image is processed to get a single line caption. The image is taken as input and it is processed and is passed through trained CNN and RNN models. At last a single line caption is obtained as output.

4. Conclusion

This paper implemented a model that produces natural language description of image regions in the form of a dataset of images and sentences, based on weak labels and with very few hard coded assumptions. We have shown that this model provides the state of the art performance in experiments on image sentence ranking.

As for attention, our model is only able to recognize the most important part of the images. That is, the attentions at each step are the same. The overall information of the image has been feed into the decoder, which is enough to generate a decent sentence, and thus the following inputs can be coarser. This is exactly the motivation of our other models. They are potential to work better given more fine tune.

Acknowledgment

The success and final outcome of this project required a lot of guidance and assistance from many people and we are extremely fortunate to have got their support all along the completion of our project.

We take this opportunity to express our profound gratitude and deep regards to our Project Guide and Project Coordinator Mr. Athmaranjan K, Associate Professor, Department of Information Science and Engineering, for his exemplary guidance, monitoring and constant encouragement through-out the course of this project. He has communicated various ideas for improving the project and has been especially enthusiastic in giving his opinions and critical reviews in a constructive manner. We will remember his contribution forever. We express a deep sense of gratitude to Mr. Janardhana Bhat K, Head of the Department, Information Science and Engineering, for his cordial support, valuable information and guidance, which helped us in completing this project through various stages.

We also express our heartfelt gratitude to our Principal Dr. Shrinivasa Mayya D, for his kind co-operation and encouragement which helped us in the completion of this project. We also thank our Management who helped us directly and indirectly for the successful completion of our project. We are thankful to and fortunate enough to get constant encouragement, support and guidance from all the Teaching and Non-teaching staff of Department of Information Science and Engineering who helped us in successfully completing our project.

References

[1] Szegedy C, Liu W, Jia Y, et al. “Going deeper with convolutions”, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

[2] Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan. “Show and tell: A neural image caption generator,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[3] Donahue, Jeffrey, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. “Long-term recurrent convolutional networks for visual recognition and description,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[4] Larochelle H, Hinton G E. "Learning to combine foveal glimpses with a third-order Boltzmann machine", Advances in neural information processing systems, 2010.

[5] Denil M, Bazzani L, Larochelle H, et al. "Learning where to attend with deep architectures for image tracking", Neural computation, 2012.

[6] Mert Kilickaya, Burak Kerim Akkus, Ruket Cakisi, "Data-Driven Image Captioning via Salient Region Discovery", IET Computer Vision Conference, China 2017.

[7] R. Kiros, R. Salakhutdinvo, Richard Zemel, "Multimodal Neural Language Models", University of Toronto, 2014.

[8] Farhad Ali, Danilids K, "Every Picture Tells a Story: Generating Sentences from Images", 2010.

[9] Alexander G, Schwing, Jyoti Aneja, "Convolutional Image Captioning", 2018.