



Article Title: **Image Caption Generation: A Review**

Image Caption Generation: A Review

Lijo B Vijay ¹, Gini R ²

¹ Student, Department of MCA, Narayanaguru College of Engineering, Anna University, Chennai.

² Assistant Professor, Department of MCA, Narayanaguru College of Engineering, Anna University.

ABSTRACT

In Artificial Intelligence (AI), the contents of an image are generated automatically which involves computer vision and NLP (Natural Language Processing). The neural model which is regenerative, is created. It depends on computer vision and machine translation. This model is used to generate natural sentences which eventually describes the image. This model consists of Convolutional Neural Network (CNN) as well as Recurrent Neural Network (RNN). The CNN is used for feature extraction from image and RNN is used for sentence generation. The model is trained in such a way that if input image is given to model it generates captions which nearly describes the image. The accuracy of model and smoothness or command of language model learns from image descriptions is tested on different datasets. These experiments show that model is frequently giving accurate descriptions for an input image. In this paper investigating various methods used in the existing work for the image caption generation recognition system.

Keywords: Deep Learning Image, Caption, Convolutional Neural Network (CNN), Recognition, Probability.

1 Introduction

The people communicate through language, whether written or spoken. They often use this language to describe the visual world around them. Images, signs are another way of communication and understanding for the physically challenged people. The generation of descriptions from the image automatically in proper sentences is a very difficult and challenging task [1], but it can help and have a great impact on visually impaired people for better understanding of the description of images on the web. A good description of an image is often said for 'Visualizing a picture in the mind'. The creation of an image in mind can play a significant role in sentence generation. Also, human can describe the image after having a quick glance at it. The progress in achieving complex goals of human recognition will be done after studying existing natural image descriptions. This task of automatically generating captions and describing the image is significantly harder than image classification and object recognition.

The description of an image must involve not only the objects in the image, but also relation between the objects with their attributes and activities shown in images. Most of the work done in visual recognition previously has concentrated to label images with already fixed



Article Title: Image Caption Generation: A Review

classes or categories leading to the large progress in this field. Eventually, vocabularies of visual concepts which are closed, makes a suitable and simple model for assumption. These concepts appear widely limited after comparing them with the tremendous amount of thinking power which human possesses. However, the natural language like English should be used to express above semantic knowledge, that is, for visual understanding language model is necessary. Model based on Neural Networks In order to generate description from an image, most of the previous attempts have suggested to combine all the current solutions of the above problem. Whereas, we will be designing a single model which takes an image as an input and is trained for producing a sequence of words where each word belongs to the dictionary that describes the image suitably. The relation between visual importance and descriptions moves to the text summarization problem in natural language processing (NLP). The important goal of text summarization is selecting or generating an abstract for document. In problem of image captioning, for any image we would like to generate a caption which will describe various features of that image.



Figure 1: Selection of Evaluation Results, Grouped by Human Beings

This paper proposes a model capable of generating novel descriptions from images. For this task, we have used Flickr 8k dataset consisting of 8000 images and five descriptions per image. The illustrates the dataset structure in which an image is having five natural language captions. In this work, we are using CNN as well as RNN. Pre-trained Convolutional Neural Network (CNN) is used for the image classification task. This network acts as an image



Article Title: Image Caption Generation: A Review

encoder. The last hidden layer is used as an input to Recurrent Neural Network (RNN). This network is a decoder which generates sentences. Sometimes, the generated sentence seems to lose track or predict wrong sentence than that of the original image content. This sentence is generated from description that is 978-1-5386-5257-2/18/\$31.00 ©2018 IEEE common in dataset and the sentence is weakly related to input image. Caption 1: A group is sitting around a snowy crevasse, Caption 2: A group of people sit atop a snowy mountain, Caption 3: A group of people sit in the snow overlooking a mountain scene, Caption 4: Five children getting ready to sled, Caption 5: Five people are sitting together in the snow.

2 Related Works

[1] Show and tell: A neural image caption generator. In 2015 Oriol Vinyal et al, demonstrated on an automatically describing the content of an image is a fundamental problem in artificial intelligence that connects computer vision and natural language processing. In this paper, we present a generative model based on a deep recurrent architecture that combines recent advances in computer vision and machine translation and that can be used to generate natural sentences describing an image. The model is trained to maximize the likelihood of the target description sentence given the training image. Experiments on several datasets show the accuracy of the model and the fluency of the language it learns solely from image descriptions. Our model is often quite accurate, which we verify both qualitatively and quantitatively. For instance, while the current state-of-the-art BLEU-1 score (the higher the better) on the Pascal dataset is 25, our approach yields 59, to be compared to human performance around 69. We also show BLEU-1 score improvements on Flickr30k, from 56 to 66, and on SBU, from 19 to 28. Lastly, on the newly released COCO dataset, we achieve a BLEU-4 of 27.7, which is the current state-of-the-art.

[2] Every picture tells a story: Generating sentences from image. In 2010, Ali Farhadi et al, demonstrated on a Humans can prepare concise descriptions of pictures, focusing on what they find important. We demonstrate that automatic methods can do so too. We describe a system that can compute a score linking an image to a sentence. This score can be used to attach a descriptive sentence to a given image, or to obtain images that illustrate a given sentence. The score is obtained by comparing an estimate of meaning obtained from the image to one obtained from the sentence. Each estimate of meaning comes from a discriminative procedure that is learned using data. We evaluate on a novel dataset consisting of human-annotated images. While our underlying estimate of meaning is impoverished, it is sufficient to produce very good quantitative results.

[3] Babytalk: Understanding and generating simple image descriptions. In 2013, Girish Kulkarni et al, demonstrated that we present a system to automatically generate natural language descriptions from images. This system consists of two parts. The first part, content planning, smooths the output of computer vision-based detection and recognition algorithms with statistics mined from large pools of visually descriptive text to determine the best content words to use to describe an image. The second step, surface realization, chooses



Article Title: Image Caption Generation: A Review

words to construct natural language sentences based on the predicted content and general statistics from natural language. We present multiple approaches for the surface realization step and evaluate each using automatic measures of similarity to human generated reference descriptions. We also collect forced choice human evaluations between descriptions from the proposed generation system and descriptions from competing approaches. The proposed system is very effective at producing relevant sentences for images. It also generates descriptions that are notably more true to the specific image content than previous work.

[4] Neural Machine Translation by Jointly Learning to Align and Translate. Neural machine translation is a recently proposed approach to machine translation. Unlike the traditional statistical machine translation, the neural machine translation aims at building a single neural network that can be jointly tuned to maximize the translation performance. The models proposed recently for neural machine translation often belong to a family of encoder-decoders and consists of an encoder that encodes a source sentence into a fixed-length vector from which a decoder generates a translation. In this paper, we conjecture that the use of a fixed-length vector is a bottleneck in improving the performance of this basic encoder-decoder architecture, and propose to extend this by allowing a model to automatically (soft-) search for parts of a source sentence that are relevant to predicting a target word, without having to form these parts as a hard segment explicitly. With this new approach, we achieve a translation performance comparable to the existing state-of-the-art phrase-based system on the task of English-to-French translation. Furthermore, qualitative analysis reveals that the (soft-) alignments found by the model agree well with our intuition.

[5] Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks. Recurrent Neural Networks can be trained to produce sequences of tokens given some input, as exemplified by recent results in machine translation and image captioning. The current approach to training them consists of maximizing the likelihood of each token in the sequence given the current (recurrent) state and the previous token. At inference, the unknown previous token is then replaced by a token generated by the model itself. This discrepancy between training and inference can yield errors that can accumulate quickly along the generated sequence. We propose a curriculum learning strategy to gently change the training process from a fully guided scheme using the true previous token, towards a less guided scheme which mostly uses the generated token instead. Experiments on several sequence prediction tasks show that this approach yields significant improvements. Moreover, it was used successfully in our winning entry to the MSCOCO image captioning challenge, 2015.

3 Conclusion

Using recurrent neural networks to predict sequences of tokens has many useful applications like machine translation and image description. However, the current approach to training them, predicting one token at a time, conditioned on the state and the previous correct token, is different from how we actually use them and thus is prone to the accumulation of errors



Article Title: Image Caption Generation: A Review

along the decision paths. In this paper, we proposed a curriculum learning approach to slowly change the training objective from an easy task, where the previous token is known, to a realistic one, where it is provided by the model itself. Experiments on several sequence prediction tasks yield performance improvements, while not incurring longer training times. Future work includes back-propagating the errors through the sampling decisions, as well as exploring better sampling strategies including conditioning on some confidence measure from the model itself.

References

1. A. Krizhevsky; I. Sutskever; G. E. Hinton, Year: 2012, “Imagenet classification with deep convolutional neural networks,” in Advances in Neural Information Processing Systems 25, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, pp. 1097–1105.
2. Kishore Papineni; Salim Roukos; Todd Ward; Wei-Jing Zhu, Year: 2002, “BLEU: a Method for Automatic Evaluation of Machine Translation”, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, pp. 311-318.
3. C. R. Ponce; W. Xiao; P. F. Schade, Year: 2019, “Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences,” Cell, Vol: 177, no: 4, pp. 999–1009.
4. Y. Miyawaki; H. Uchida; O. Yamashita, Year: 2008, “Visual image reconstruction from human brain activity using a combination of multiscale local image decoders,” Neuron, Vol: 60, no: 5, pp. 915–929.
5. G. Shen; T. Horikawa; K. Majima, Year: 2019, “Deep image reconstruction from human brain activity,” PLOS Computational Biology, Vol: 15, no: 1, pp. 1–23.
6. A. Krizhevsky; I. Sutskever; G. E. Hinton, Year: 2012, “Imagenet classification with deep convolutional neural networks,” in Proceedings of Advances in neural information processing systems, pp. 1097–1105.
7. S. Amirian; Z. Wang; T. R. Taha; H. R. Arabnia, Year: 2018, “Dissection of deep learning with applications in image recognition,” in Proc. Int. Conf. Comput. Sci. Comput. Intell. (CSCI), pp. 1132–1138.
8. Q. You; H. Jin; Z. Wang; C. Fang; J. Luo, Year: 2016, “Image captioning with semantic attention,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 4651–4659.
9. J. Aneja; A. Deshpande; A. G. Schwing, Year: 2018, “Convolutional image captioning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit, pp. 5561–5570.



Article Title: Image Caption Generation: A Review

10. T. Lin; M. Maire; S. Belongie; J. Hays; P. Perona; D. Ramanan; P. Dollár; C. L. Zitnick, Year: 2014, “Microsoft COCO: Common objects in context,” in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, pp. 740–755.
11. T. Iqbal; H. Ali, Year: 2018, “Generative adversarial network for medical images (MI-GAN),” J. Med. Syst., Vol: 42, no: 11, p. 231.
12. S. Amirian; K. Rasheed; T. R. Taha; H. R. Arabnia, Year: 2019, “Image captioning with generative adversarial network,” in Proc. Int. Conf. Comput. Sci. Comput. Intell. (CSCI), pp. 272–275.