Image Caption Generator Using Reinforcement Learning

Shubham Yadav
Department of CSE
ABES Institute of Technology
Ghaziabad, Uttar Pradesh
subhamyadav580@gmail.com

Ujjawal Srivastava
Department of CSE
ABES Institute of Technology
Ghaziabad, Uttar Pradesh
ujjawalk10srivastava@gmail.com

Surya Pratap
Department of CSE
ABES Institute of Technology
Ghaziabad, Uttar Pradesh
pratap.surya7799@gmail.com

ABSTRACT

Image captioning is the task of generating textual descriptions of a given image, requiring techniques of computer vision and natural language processing. Recent models have utilized deep learning techniques to this task to gain performance improvement. However, these models can neither distinguish more important objects than others in a given image, nor explain the reasons why specific words have been selected when generating captions. To overcome these limitations, this paper proposes an explainable image captioning model, which generates a caption by indicating specific objects in a given image and providing visual explanation using them. The model has been evaluated with datasets such as MSCOCO, Flickr8K, and Flickr30K, and some qualitative results are presented to show the effectiveness of the proposed model. 8

1. INTRODUCTION

In the past few years, computer vision in image processing area has made significant progress, like image classification [1] and object detection [2]. Benefiting from the advances of image classification and object detection, it becomes possible to automatically generate one or more sentences to understand the visual content of an image, which is the problem known as Image Captioning. Generating complete and natural image descriptions automatically has large potential effects, such as titles attached to news images, descriptions associated with medical images, text-based image retrieval, information accessed for blind users, human-robot interaction. These applications in image captioning have important theoretical and practical research value. Therefore, image captioning is a more complicated but meaningful task in the age of artificial intelligence.

Given a new image, an image captioning algorithm should output a description about this image at a semantic level. For example, in Fig. 1, the input image consists of people, boards and the waves. In the bottom, there is a sentence describing the content of the image— the objects emerging in the image, the action and the scene are all described in this sentence. For the image captioning task, humans can easily understand the image content and express it in the form of natural language sentences according to specific needs; however, for computers, it requires the integrated use of image processing, computer vision, natural language processing and other major areas of research results. The challenge of image captioning is to design a model that can fully use image information to generate more human-like rich image descriptions. The meaningful description generation process of high level image semantics requires not only the understanding of objects or scene recognition in the image, but also the ability to analyse their states, understand the relationship among them and generate a semantically and syntactically correct sentence. It is currently unclear how the brain understands an image and organizes the visual information into a caption. Image captioning involves a deep understanding of the world and which things are salient parts of the whole.

Fig. 1. A brown dog in the snow

Provided an image, the motive of image captioning is to form a sentence that is grammatically credible and semantically valid to the content of the image as shown in Figure 1. This process involves two steps: Visual processing and linguistic processing. To assure that the generated captions are grammatically and semantically correct and to deal with problems arising from the corresponding modality and integrated competently, techniques of computer vision and NLP are utilized So, by this end, many methods discussed below. Though the image captioning task is complex, the latest breakthroughs in deep neural networks [11–16], used extensively in the domain of computer vision [17–20] and NLP [21–24], made it easier and hence image caption generating machines based on deep neural networks came into existence. Robust deep neural networks implement effective solutions to visual and language modelling. Therefore, they are used to supplement existing systems and design many new approaches. Engaging deep neural networks to handle the image captioning task resulted in state-of-the-art outcomes [25–30]. With the recent progress in transfer learning and image captioning, we propose a novel architecture that compares multiple transfer learning models based on different metrics includes BLEU Score and others.

2. CNN-RNN BASED FRAMEWORK

In human's eyes, an image consists of different colours to compose the different scenes. But in the view of computer, most images are painted with pixels in three channels. However, in the neural network, different modalities of data are all trending to create a vector and do the following operations on these features. It has been convincingly shown that CNNs can produce a rich representation of the input image by embedding it into a fixed-length vector, such that this representation can be used for a variety of vision tasks like object recognition, detection and segmentation [8]. Hence, image captioning methods based on encoder-decoder frameworks often use a CNN as an image encoder. The RNN network obtains historical information through continuous circulation of the hidden layer, which has better training capabilities and can perform better than mining deeper linguistic knowledge such as semantics and syntax information implicit in the word sequence [9]. For a dependency relationship between different location words in historical information, a recurrent neural network can be easily represented in the hidden layer state. In image captioning task based on encoder-decoder framework, the encoder part is a CNN model for extracting image features. It can use models such as AlexNet [1], VGG [10], GoogleNet [11] and ResNet [12]. In the decoder part, the framework enters the word vector expression into the RNN model. For each word, it is first represented by a one-hot vector, and then through the word embedding model, it becomes the same dimension as the image feature. The image captioning problem can be defined in the form of a binary(I, S), where I represents a graph and S is a sequence of target words, S = S 1 , S 2 and S i is a word from the data set extraction. The goal of training is to maximize the likelihood estimation of the target description p(S—I) for the goal of the generated statement and the target statement matching more closely. Mao et al. [13] proposed a multimodal Recurrent Neural Network(m-RNN) model that creatively combines the CNN and RNN model to solve the image captioning problem. Because of the gradient disappearance and the limited memory problem of ordinary RNN, the LSTM model is a special type of structure of the RNN model that can solve the above problems. It adds three control units (cell), which are the input, output and forgot gates. As the information enters the model, the information will be judged by the cells. Information that meets the rules will be left, and nonconforming information will be forgotten. In this principle, the long sequence dependency problem in the neural network can be solved. Vinyals et al. [14] proposed the NIC (Neural Image Caption) model that takes an image as input in the encoder part and generates the corresponding descriptions with LSTM networks in the decoder part. The model solves the problem of vectorization of natural language sentences very well. It is of great significance to use computers dealing with natural language, which makes the processing of computers no longer stays at the simple level of matching, but further to the level of semantic understanding. Inspired by the neural network-based machine translation framework, the attention mechanism in the field of computer vision is proposed to promote the alignment between words and image blocks. Thereby, in the process of sentence generation, the "attention" transfer process of simulating human vision can be mutually promoted with the generation process of the word sequence, so that the generated sentence is more in line with the people's expression habit. Instead of encoding the whole image as a static vector, the attention mechanism adds the whole and spatial information corresponding to the image to the extraction of the image features, resulting in a richer statement description. At this time, the image features are considered as the dynamic feature vectors combined with the weights information. The first attention mechanism was proposed in [15], it proposed the "soft attention" which means to select regions based on different weights and the "hard attention" which performs attention on a particular visual concept. The experimental results obtained by using attention-based deep neural networks have achieved remarkable results. Using attention mechanism makes the model generate each word according to the corresponding region of an image as is shown in Fig.

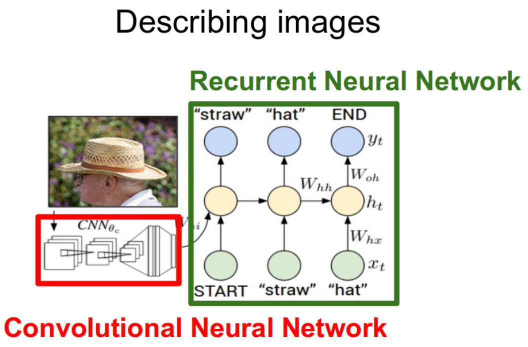

Fig. 2. The image above shows the image based model.

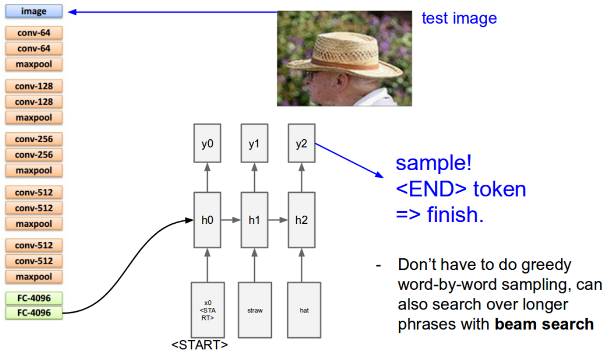

Fig. 3. The image above shows the language based model.

However, it also suffers from two main drawbacks for the image captioning task, which motivate further significant research. The first is that the metrics used for testing and loss for training are different. We use cross- entropy as loss, but metrics are non-differentiable and cannot be directly used as training loss. And log likelihood can be seen as giving the same weight to each word, but in fact people evaluate different words with selective weights. This discrepancy is known as “loss-evaluation mismatch” problem. The second is that when training, the input of each time step comes from the real caption and when generated, each word generated is based on the previously generated word; Once a word is not generated well, it may get far away from the ground truth. This discrepancy is known as “exposure bias” problem.

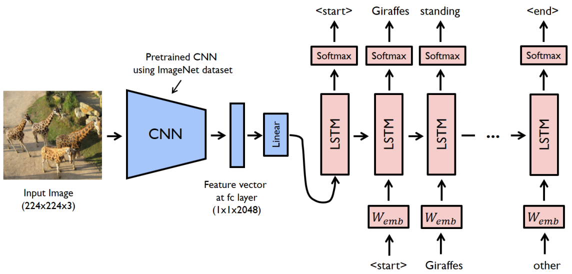

Fig. 4. The image above shows the walkthrough, a pre-trained resnet-152 model is used as an encoder, and the decoder is an LSTM network.

3. REINFORCEMENT BASED FRAMEWORK

Reinforcement learning has been widely used in gaming, control theory, etc. The problems in control or gaming have concrete targets to optimize by nature, whereas defining an appropriate optimization goal is nontrivial for image captioning. When applying the reinforcement learning into image captioning, the generative model (RNN) can be viewed as an agent, which interacts with the external environment (the words and the context vector as the input at every time step) . The parameters of this agent define a policy, whose execution results in the agent picking an action. In the sequence generation setting, an action refers to predicting the next word in the sequence at each time step. After taking an action the agent updates its internal state (the hidden units of RNN) . Once the agent has reached the end of a sequence, it observes a reward. In such a framework, the RNN decoder acts like a stochastic policy, where choosing an action corresponds to generating the next word. During training PG method chooses actions according to the current policy and only observe a reward at the end of the sequence (or after maximum sequence length) , by comparing the sequence of actions from the current policy against the optimal action sequence. The goal of training is to find the parameters of the agent that maximize the expected reward. The idea of using PG (policy gradient) to optimize non differentiable objectives for image captioning was first proposed in the MIXER paper [21], by treating the score of a candidate sentence as analogous to a reward signal in a reinforcement learning setting. In the MIXER method, since the problem setting of text generation has a very large action space which makes the problem be difficult to learn with an initial random policy, it takes actions of training the RNN with the cross-entropy loss for several epochs using the ground truth sequences which makes the model can focus on a good part of the search space. This is a new form of training that mixes together the MLE (maximum likelihood estimation) and the reinforcement objective. This reinforcement learning model is driven by visual semantic embedding, which performs well across different evaluation metrics without re-training. Visual- semantic embedding, which provides a measure of similarity between images and sentences, can measure similarities between images and sentences, the correctness of generated captions and serve a reasonable global target to optimize for image captioning in reinforcement learning. Instead of learning the sequential loop model to greedily find the next correct word, the decision-making network uses the “policy network” and the “value network” to jointly determine the next best word for each time step. The policy network provides the confidence of predicting the next word according to current state. The value network evaluates the reward value of all possible extensions of the current state.

Table 1. Training time for one minibatch
on Flicker dataset

Method	Parameter	Time/Epoch
CNN-RNN	13M	1529S
CNN-CNN	19M	1585s
Reinforcement	14M	3930s

In Table 1, we compare the training parameters and training time (in seconds) for RNN, CNN and Reinforcement Framework. The timings are obtained on Nvidia Geforce 940MX GPU. We train a CNN faster per parameter than the RNN and Reinforcement framework. But as for the accuracy and the diversity, the performance of CNN is worse than the other models, which is illustrated in the following section.

4. EVALUATION METRICS

The current study mostly uses the degree of matching between the caption sentence and the reference sentence to evaluate the pros and cons of the generation results. The commonly used methods include BLEU [16], METEOR [17], ROUGE [18], CIDEr [19], and SPICE [20] these five measurement indicators. Among them, BLEU and METEOR are derived from machine translation, ROUGE is derived from text abstraction, and CIDEr and SPICE are specific indicators based on image captioning.

—BLEU is widely used in the evaluation of image annotation results, which is based on the n-gram precision. The principle of the BLEU measure is to calculate the distance between the evaluated and the reference sentences. BLEU method tends to give the higher score when the caption is closest to the length of the reference statement.

—ROUGE is an automatic evaluation standard designed to evaluate text summarization algorithms. There are three evaluation criteria, ROUGE-N, ROUGE-L, and ROUGE-S. ROUGE-N is based on the given sentence to be evaluated, which calculates a simple n-tuple recall for all reference statements: ROUGE-L is based on the largest common sequence (LCS) calculating the recall. ROUGE-S calculates recall based on co- occurrence statistics of skip-bigram between reference text description and prediction text description.

—CIDEr is the special method which is provided for the image captioning work. It measures consensus in image captioning by performing a term frequency inverse document frequency (tf-idf) for each n-gram. Studies have shown that the match between CIDEr and human consensus is better than other evaluation criteria.

—METEOR is based on the harmonic mean of unigram precision and recall, but the weight of the recall is higher than the accuracy. It is highly relevant to human judgment and differs from the BLEU in that it is not only in the entire set, but also in the sentence and segmentation levels, and it has a high correlation with human judgment.

—SPICE evaluates the quality of image captions by converting the generated description sentences and reference sentences into graph-based semantic representations, namely “scene graphs”. The scene graphs extract lexical and syntactic information in natural language and explicitly represents the objects, attributes, and relationships contained in the image.

5. CONCLUSION

Image captioning has made significant advances in recent years. Recent work based on deep learning techniques has resulted in a breakthrough in the accuracy of image captioning. The text description of the image can improve the content-based image retrieval efficiency, the expanding application scope of visual understanding in the fields of medicine, security, military and other fields, which has a broad application prospect. At the same time, the theoretical framework and research methods of image captioning can promote the development of the theory and application of image annotation and visual question answering (VQA), cross media retrieval, video captioning and video dialog, which has important academic and practical application value.

6. REFERENCES

—Haoran Wang , Yue Zhang and Xiaosheng Yu ”An Overview of Image Caption Generation Methods” Hindawi (January 2020)

—Ali Mollaahmadi Dehaqi , Vahid Seydi and Yeganeh Madadi ”Adversarial Image Caption Generator Network” Springer (MAY 2021)

—Megha J Paniker , Vrinda Mathur and Vikas Upadhayay ”Image Caption Generator” IJITEE (January 2021)

—Chetan Amritkar and Vaishali Jabade ”Image Caption Generation Using Deep Learning Technique” IEEE (25 April 2019)

—Haoran Wang , Yue Zhang, and Xiaosheng Yu ”An Overview of Image Caption Generation Methods” Hindawi Computational Intelligence and Neuroscience Volume 2020, Article ID 3062706, 13 pages

—Phyu Phyu Khaing ”Two-Tier LSTM Model for Image Caption Generation” Article in International Journal of Intelligent Engineering and Systems · June 2021

—Shuang Liu , Liang Bai , Yanli Hu and Haoran Wang ”Image Captioning Based on Deep Neural Networks” MATEC Web of Conferences 232, 01052 (2018) EITCE 2018

—Haichao Shi, Peng Li, Bo Wang and Zhenyu Wang ”Image Captioning based on Deep Reinforcement Learning” ICIMCS’18, August 17-19, 2018, Nanjing, China

—S. -H. Han and H. -J. Choi, ”Explainable Image Caption Generator Using Attention and Bayesian Inference,” 2018 International Conference on Computational Science and Computational Intelligence (CSCI), 2018, pp. 478-481, doi: 10.1109/CSCI46756.2018.00098.

—A. Hani, N. Tagougui and M. Kherallah, ”Image Caption Generation Using A Deep Architecture,” 2019 International Arab Conference on Information Technology (ACIT), 2019, pp. 246-251, doi: 10.1109/ACIT47987.2019.8990998.

—Gu, Jiuxiang, et al. ”Stack-Captioning: Coarse-to-Fine Learning for Image Captioning.” (2018)

—Krizhevsky, Alex, I. Sutskever, and G. E. Hinton. ”ImageNet classification with deep convolutional neural networks.” International Conference on Neural Information Processing Systems Curran Associates Inc. 1097-1105. (2012)