Understanding Image Caption Algorithms: A Review

Cao Chenyu

Hangzhou Dianzi University, Xiasha District, Hangzhou City, Zhejiang Province

chcao@hdu.edu.cn

Abstract. The technology of Image caption is developing rapidly. In order to review the recent advancement in this field, this article briefly summarize several typical works in image caption researching, in which they figured out new ways to improve the accuracy or efficiency. We describe the methods, organize the results of experiments in one form, and then analyse the data. Besides, a novel quantitative metric which can measure the quality of image caption more objectively is also introduced.

1. Introduction

Recently, Image caption has become a major branch of computer vision and deep learning. It's a technology that uses algorithms to generate descriptive language to an image. The advancement of this technology is aligned with the development of artificial intelligence. Due to its immaturity, this technology has not been widely used at present stage. But its potential is limitless and waited to be exploited. A reasonable inference is that it can free the person who interprets the image from work, and it can also be used to help the blind to understand surrounding environment, or in many fields of automation like baggage security check, auto-surveillance and unmanned vehicle. So this subject, along with natural language processing and machine translation, is one of the most important products of artificial intelligence and has highly research value.

The research towards image caption could be divided into two categories. One concentrates on the problem itself. They improve the existing algorithms, using various mechanisms to train models and creating a new joint model. And many of the articles have drawn their inspiration from further imitating the human brain. For instance, Xu K, et al. [1] mimics the attention in human visual system and introduce an attention-based model. Moreover, they presents two variants: a "hard" stochastic attention mechanism and a "soft" deterministic attention mechanism. Vinyals O, et al. [2] presents a generative model based on a deep recurrent architecture, it's a joint model that compromises not only visual understanding but also a language model that does the generation of caption. Chen X, et al. [3] proposes a bi-directional representation that can generate both novel descriptions from images and visual representations from descriptions, and it can also dynamically captures the visual aspects of the scene that have already been described. Jin J, et al. [4] proposes an image caption system that exploits the parallel structures between images and sentences. They hypothesis that visual perception and the order of words generation in a sentence is highly correlated, so they encode what is semantically shared by both the visual scene and the text description. Moreover, they have achieved identifying the environment in the image and generate sentences that match the environment to improve the accuracy.

The other concentrates on analysing the demerits of the current metrics and propose a new, more accurate metric. This branch is less studied, but is also vital. A typical example is the work presented by Anderson P, et al.[5] In this article, the authors analyse the current metrics' demerits using specific

examples and conclude that the current metrics that are primarily sensitive to n-gram overlap have limitations regarding to score the captions that have similar sentence structure. Then they presents a new metric called "SPICE", it compares the semantic propositional content to generate scores, and this metric better grasps the human judgment of the image caption.

The analysis in this article covers several state-of-the-art articles of Image Caption. This article systematically analyses the core concept of each paper, points out the chronological order, logic order, advantages and disadvantages of these articles, and presents the development course of study of image captions. Moreover, experiments are conducted to verify the analysis.

The main contribution of this article is summarizing the current state and pointing out the future trends of this field.

2. Methods

As a good example of getting inspiration from human brain, Xu K, et al. [1] studies the existence of attention in human vision system, and introduces a image caption model based on "attention". In this model, the attention means the ability to dynamically emphasizes the salient object in an image. Specifically, Xu K, et al. [1] show how the model can automatically adjust its gaze on the important object when generating the corresponding words. To achieve this, they come up with two mechanisms: a "hard" stochastic attention mechanism and a "soft" deterministic attention mechanism, and trains them by standard back-propagation methods and maximizing an approximate variational lower bound or equivalently. Moreover, this model has an advantage that it can approximately visualizing what the model "sees" to gain insights. Yet a potential shortcoming is that the focus on the most salient object will cause the loss of the other less important information, and may result in the less complete and less abundant captions. As an example of understanding the principle, the attention mechanism is also been widely used in machine translation, the diagram is shown below in Figure 1.

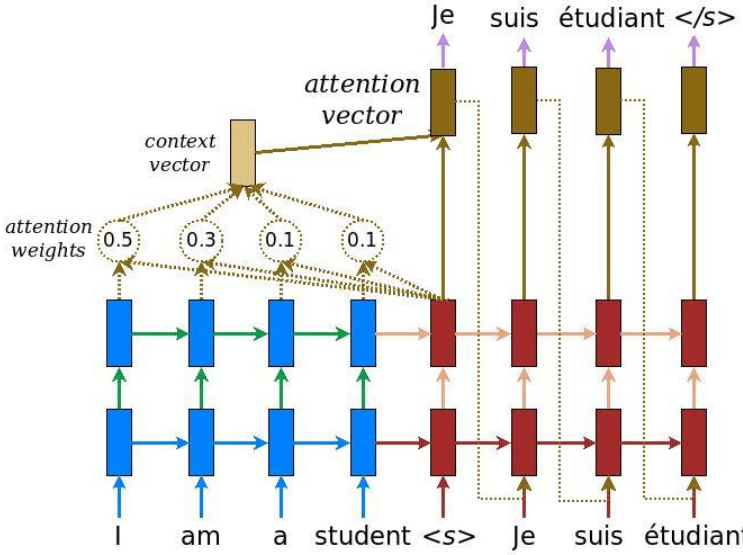

Figure 1. The attention mechanism in machine translation [6]

As is known to us all, image caption requires two steps: to understand the image and to generate words. And the latter needs a language model. Regarding to this problem, many researchers solve the two procedures independently and then splice them. However, this is not efficient enough as the brain compress large amount of vision information into descriptive language in a very short time. Impressed by the latest progress in machine translation that recursive neural networks (RNN) can accomplish the translation work that previously required a series of independent tasks, and even in a much simpler and more accurate way, Vinyals O, et al. [2] proposes a generative model based on a deep recurrent architecture. The deep convolutional neural network (CNN) is used instead of the encoder RNN, which

is first pre-trained for an image classification task, and then the last hidden layer is used as the input to the RNN decoder that generates the sentence. This is an end-to-end system, a fully trainable neural network which can be optimized by stochastic gradient descent. The objective function of such model is maximizing the conditional likelihood p(S|I), where I is the input image and S is the generated sentence. They adopt a Long-Short Term Memory-based Sentence Generator, which is widely used in translation and generation tasks, as their RNN option. Schematic diagram of memory block is shown in Figure 2.

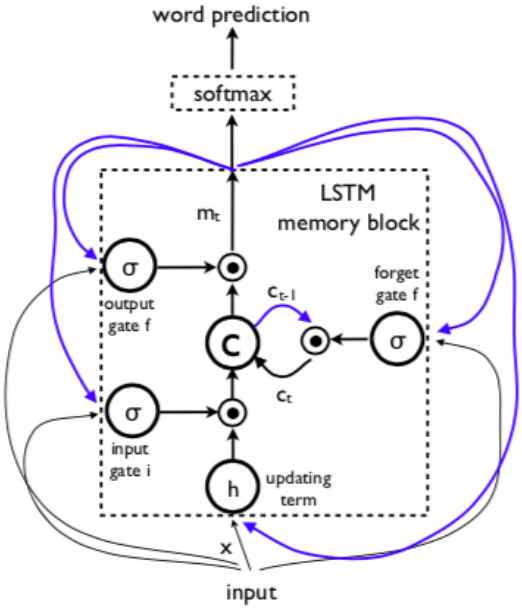

Figure 2. The memory block contains a cell which is controlled by 3 gates [2]

Beside these two achievements, Chen X, et al. [3]'s paper that based on the further imitation towards human brain was published in the same year at 2015. This work was inspired by the facts that human always create a mental image naturally when they understand sentences, and that the image in the brain tends to stay longer than the precisely sentences. Applying this into models, Chen X, et al. [3] discussed the study of images and their captions' joint feature spaces. They project image features and sentence features into a common space, and both the new descriptions could be generated from the images and the vision representation could be generated from descriptions. The model can also dynamically captures the vision scenes from the descripted images, i.e. when a word is generated or read, the vision representation will be renewed to reflect the new information. This process resembles the long-term memory of the concepts. The article uses RNNs to accomplish this. Moreover, the article also includes comparisons between final models and RNN baseline, and the accuracy of bi-directional retrieval is examined. Schematic diagram of hidden unit is shown in Figure 3.

In human visual perception, there is a thread of visual order when the attentional shift between parts of an image, which is explained as the order that human brain understands a series of abstract meanings implied by the image when observing it. Based on this hypothesis, Jin J, et al. [4] proposes a generation model that utilizes the parallel structure between images and sentences. Specifically, the article assumes that there is a close correspondence between visual concepts and textual realizations, and that the process of generating the next word based on previously generated words is consistent with the human visual perception experience. To achieve this, the article encodes the semantically shared content between visual scenes and text description, and uses recursive neural RNNs to build the model. The hidden state of this network is used to predict where the next visual focus should be, and to determine what the next word in the corresponding text should be. The article also introduces another model concerning the scenario-specific contexts. The model captures high-level semantic information encoded in the image,

such as the effect of the location at which the image was taken on the possible activity of the people in the image. They trained the language model to generate words that agree with the specific scene types. To achieve this, the scene context extracts visual feature vectors from the image, affecting the generation of words by biasing the parameters in the RNNs. Schematic diagram is shown in Figure 4.

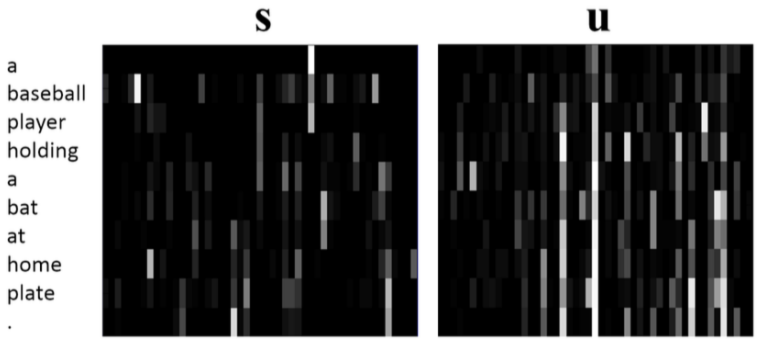

Figure 3. The hidden units s and u activations through time (vertical axis) [3]

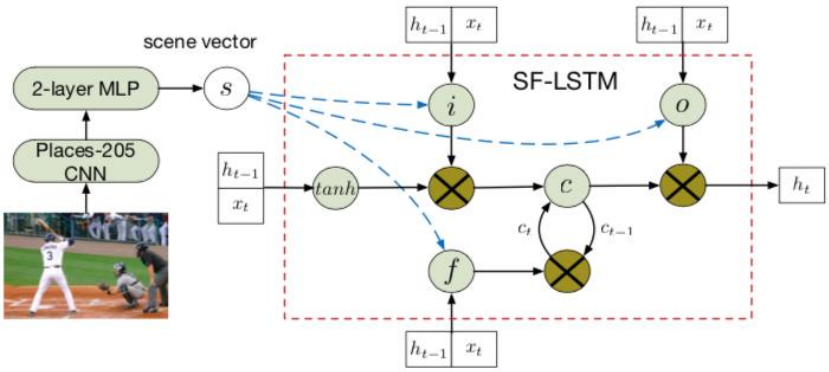

Figure 4. The right part is the basic unit of LSTM. s is used to factor the weight matrix in the 3 gates. [4]

The existing evaluation metrics of captions mainly judge the similarity of the descriptions of natural language by testing their sensitivity towards n-gram overlap. However, these methods have a strong limitation that they would give high scores to sentences that just have the same structure, and it deviates from human perception of an image in many cases. Based on this fact, Anderson P, et al.[5] propose a new caption evaluation metric called SPICE. This metric gives evaluations by analyzing the semantic content of the description, better grasps the human judgement to images. Moreover, SPICE could analyze the performance of any model more detailed than other automated evaluation indicators. For instance, it can analyzes which model can best express color of the image and whether a model can count. To achieve these functions, they use a dependency parser pre-trained on large dataset to establish the syntactic dependencies between words, and then map from dependency trees to scene graphs using a rule-based system. For the provided candidate and reference scenes, SPICE computes an F-score which defined over the conjunction of logical tuples representing semantic propositions in the scene graph. Diagram of the efficiency of SPICE compared with other metrics is shown in Figure 5.

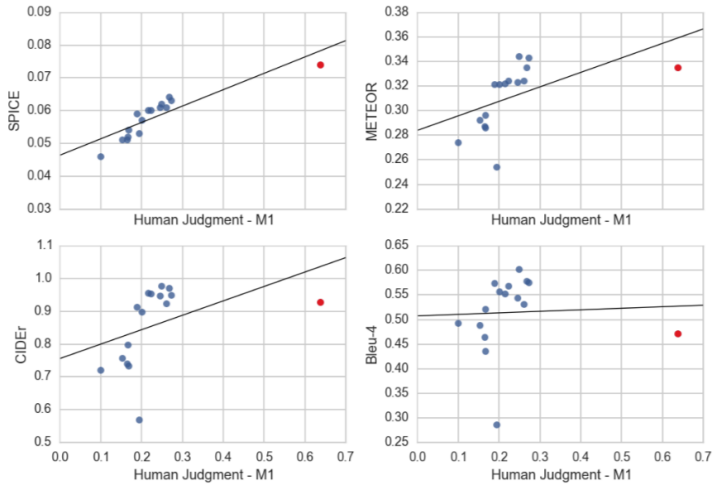

Figure 5. Evaluation metrics vs. human judgments for the 15 entries. Each data point represents a single model, and captions produced by human are marked in red. [5]

3. Experiments

Table 1. Scores

dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDER
Flickr8k	Soft-Attention(Xu et al., 2015) [1]	67	44.8	29.9	19.5	18.93	
	Hard-Attention(Xu et al., 2015) [1]	67	45.7	31.4	21.3	20.3	
	Neural Image Caption(Vinyals et al., 2015) [2]	63					
	RNN model(Chen et al., 2015) [3]		14.1			17.97	
Flickr30k	Soft-Attention(Xu et al., 2015) [1]	66.7	43.4	28.8	19.1	18.49	
	Hard-Attention(Xu et al., 2015) [1]	66.9	43.9	29.6	19.9	18.46	
	Neural Image Caption(Vinyals et al., 2015) [2]	66					
	RNN model(Chen et al., 2015) [3]		12.6			16.42	
MS COCO	Soft-Attention(Xu et al., 2015) [1]	70.7	49.2	34.4	24.3	23.9	
	Hard-Attention(Xu et al., 2015) [1]	71.8	50.4	35.7	25	23.04	
	Neural Image Caption(Vinyals et al., 2015) [2]				27.7	23.7	85.5
	RNN model(Chen et al., 2015) [3]		18.35			20.04	
	RNN model + FT(Chen et al., 2015) [3]		18.99			20.42	

Pascal	RNN-(RA+SF)-BEAM(Jin et al., 2015) [4]	69.7	51.9	38.1	28.2	23.5	83.8
	Neural Image Caption(Vinyals et al., 2015) [2]	59					
	RNN model(Chen et al., 2015) [3]			10.48		16.69	
	RNN model + FT(Chen et al., 2015) [3]			10.77		16.87	

In the experiments, we induct the results of the 4 experiments conducted by 4 papers introduced above , list the most representative data into one sheet to compare and analyze the results. But at first, we would look into the datasets and the metrics they choose.

As we can see from Table 1, among the four articles, 3 of them use Flickr8k and Flickr30k, all of them use MS COCO, and 2 of them use Pascal. These datasets contain 8000, 31000, 123,000, and 9963 images respectively and sentences in English describing these images. Specifically, the Flickr dataset is the benchmark for image captions. It was published by Yahoo, almost every image in it has been annotated by labelers with 5 sentences that are relatively visual and unbiased. The MS COCO is released by Microsoft, and it is a large-scale object detection, segmentation, and captioning dataset. And in order to maintain the same number of references among the datasets, some articles state that they discard caption in excess of 5. The PASCAL dataset is customary used for testing only after a system has been trained on different dataset such as any of the other 3 datasets, two of its main folders are the Annotation folder and the ImageSets folder. The former stores the xml file, which is an explanation of the image. Each image is for an xml file with the same name; the latter holds txt files, which divide the images of the dataset into various collections.

All articles use BLEU and METEOR as metrics. BLEU adopts an N-gram matching rule, and the principle of it is to compare the similarity of n groups of words between the translation and the reference translation. Unigram precision $P = \frac{m}{w_t}$, where m is number of words from the candidate that are found in the reference, and w_t is the total number of words in the candidate[7]. The METEOR metric is based on the harmonic mean of unigram precision and recall, with recall weighted higher than precision. It can solve some of the defects inherent in the BLEU metric. Unigram precision P is computed as: $P = \frac{m}{w_t}$. Unigram recall R is calculated as: $R = \frac{m}{w_r}$, where w_r is the number of unigrams in the reference translation. In addition, the harmonic mean combines precision and recall, and the weight of recall is 9 times that of precision: $F_{mean} = \frac{10PR}{R+9P}$[8]. CIDEr treats each sentence as a "document", expresses it as a form of tf-idf vector, and then calculates the cosine similarity of the reference caption and the caption generated by the model.

As is generalized above, all methods behave well in the listed datasets. And their scores are also similar. This results illustrate the high efficiency of each model. Relatively, the Hard-Attention(Xu et al., 2015) [1] behaves best in the dataset of Flickr8k and Flickr30k. RNN-(RA+SF)-BEAM(Jin et al., 2015) [4] behaves best in the dataset of MS COCO and the metric of BLEU-2, BLEU-3, and BLEU-4. Neural Image Caption (Vinyals et al., 2015) [2] behaves best in the dataset of MS COCO and the metric of METEOR and CIDER. Most of the metrics have higher BLEU-1 and BLEU-2 value than BLEU-3 and BLEU-4, indicating that these metrics behaves better in short sentences than long sentences, which is one of its limitation. But in general, all methods demonstrate the state-of-art performance in image caption.

4. Conclusion

This paper reviews several high efficient image caption methods by introducing their principles, algorithms and experimental performances. As we can see, improvements considering various aspects of image caption are reached as listed above. Although image caption is not widely applied at present

stage, it is developing at full speed. When the day comes that this technology is mature, it will be a huge step in promoting computer's intelligence. And as increasingly number of researches are conducted on this topic, this future technology will become closer to reality.

References

[1] Xu K, Ba J, Kiros R, et al. 2015 Show, attend and tell: Neural image caption generation with visual attention *International conference on machine learning*. 2048-2057.
[2] Vinyals O, Toshev A, Bengio S, et al. 2015 Show and tell: A neural image caption generator *Proceedings of the IEEE conference on computer vision and pattern recognition.* 3156-3164
[3] Chen X, Lawrence Zitnick C. 2015 Mind's eye: A recurrent visual representation for image caption generation *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2422-2431.
[4] Jin J, Fu K, Cui R, et al. 2015 Aligning where to see and what to tell: image caption with region-based attention and scene factorization. arXiv preprint arXiv:1506.06272.
[5] Anderson P, Fernando B, Johnson M, et al. 2016 Spice: Semantic propositional image caption evaluation *European Conference on Computer Vision*. Springer, Cham, 382-398.
[6] A Brief Overview of Attention Mechanism https://medium.com/syncedreview/a-brief-overview-of-attention-mechanism-13c578ba9129
[7] Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. J. 2002. BLEU: a method for automatic evaluation of machine translation. ACL-2002: *40th Annual meeting of the Association for Computational Linguistics.* pp. 311–318. CiteSeerX 10.1.1.19.9416
[8] Banerjee, S. and Lavie, A. (2005) "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments" in *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd*