# Image Captioning Based on Deep Neural Networks

Shuang Liu[1], Liang Bai[1,a], Yanli Hu[1] and Haoran Wang[1]

[1]*College of Systems Engineering, National University of Defense Technology,410073 Changsha, China*

**Abstract.** With the development of deep learning, the combination of computer vision and natural language process has aroused great attention in the past few years. Image captioning is a representative of this filed, which makes the computer learn to use one or more sentences to understand the visual content of an image. The meaningful description generation process of high level image semantics requires not only the recognition of the object and the scene, but the ability of analyzing the state, the attributes and the relationship among these objects. Though image captioning is a complicated and difficult task, a lot of researchers have achieved significant improvements. In this paper, we mainly describe three image captioning methods using the deep neural networks: CNN-RNN based, CNN-CNN based and Reinforcement-based framework. Then we introduce the representative work of these three top methods respectively, describe the evaluation metrics and summarize the benefits and major challenges.

## 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.

A couple of people riding waves on top of boards.

**Figure 1** An example of image captioning

Despite such challenges, the problem has achieved significant improvements over the past few years. Image captioning algorithms are typically divided into three categories. The first category, as shown in Fig2. (a), tackles this problem using the retrieval-based methods, which first retrieves the closest matching images, and then transfer their descriptions as the captions of the query

[a] Corresponding author: xabpz@163.com

images [3]. These methods can produce grammatically correct sentences but cannot adjust the captions according to the new image. The second category in Fig2. (b), typically uses template-based methods to generate descriptions with predefined syntactic rules and slit sentences into several parts [4]. These methods first take advantage of several classifiers to recognize the objects, as well as their attributes and relationships in an image, and then use a rigid sentence template to form a complete sentence. Though it can generate a new sentence, these methods either cannot express the visual context correctly or generate flexible and meaningful sentences.

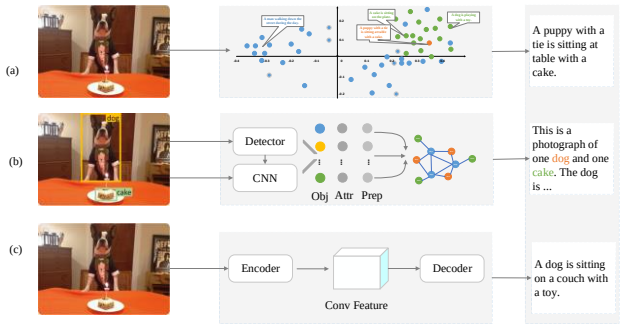


**Figure 2** Three catogories for image captioning

With the extensive application of deep learning, most recent works fall into the third category called neural network-based methods in Fig2. (c). Inspired by machine learning's encoder-decoder architecture [5], recent years most image captioning methods employ a Convolutional Neural Network (CNN) as the encoder and a Recurrent Neural. Network (RNN) as the decoder, especially Long Short-Term Memory (LSTM) [6] to generate captions [7], with the objective to maximize the likelihood of a sentence given the visual features of an image. Some methods are using CNN as the decoder and the reinforcement learning as the decision-making network.

According to these different encoding and decoding methods, in this paper, we divide the image captioning methods with neural networks into three categories: CNN-RNN based, CNN-CNN based and reinforcement-based framework for image captioning. In the next part, we will talk about their main ideas.

## 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, $S = \{S_1, S_2 \cdots\}$ 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.3.

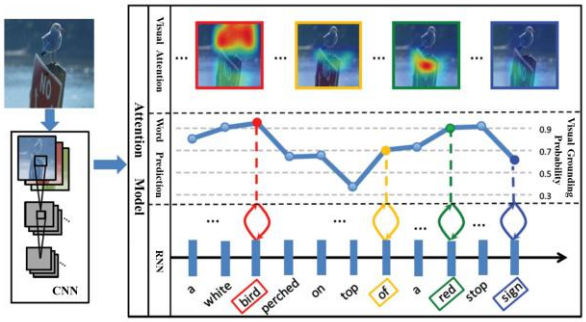


**Figure 3** Illustration of the Attention Model

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 [21]. 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 [21].

## 3 CNN-CNN based framework

Although models like LSTM networks have memory cells which can memorize the long history information of the sequence generation process better than RNN, it is still updated at each time, which render the long-term memory rather difficult. Inspired by the work of machine learning, recent works have shown benefits of CNN on the image captioning work. Using the CNN in NLP for text generation has been proved very powerful [22]. In the field of neural machine translation, it has proved that the CNN convolution model is used to replace the RNN recurrent model, which not only exceeds the accuracy of the cycle model, but also increases the training speed by a factor of nine. Most image captioning works are inspired by the machine translation, since the translation work is in the sequence to sequence architecture and in the image captioning task, an image is viewed as a sentence in a source language. To the best of our knowledge, the first convolutional network for the text generation process in image captioning is the work done by Aneja et al. [23] and we call this CNN-CNN based framework.

This framework contains three main components similar to the RNN technique. The first and the last components are word embeddings in both cases. However, while the centre component contains LSTM or GRU (Gated Reccurent Unit) units in the RNN case, masked convolutions are employed in the CNN-based approach. This component, unlike the RNN, is feed-forward without any recurrent function. Aneja et al. [23] has demonstrated the CNN-CNN framework has a faster training time per number of parameters but the loss is higher for CNN than RNN. The reason of the CNN model's accuracy is that CNN are being penalized for producing less-peaky word probability distributions. However, less peaky distributions are not necessarily bad, where multiple word predictions are possible for predicting diverse captions, which is shown in Fig.4.


**CNN-RNN**: A parking meter with a sign on it.
**CNN-CNN**: A doll is sitting next to a parking meter.
**Ground Truth**: A doll with articulated joints stares form her perch between two parking meters.

**Figure 4** The generated descriptions of CNN-RNN and CNN-CNN models

Actually, the layered abstraction of convolution and the triple gate of recurrence play the common role. Although the means are different, the purpose is to ignore minor content and highlight the important content. Therefore, in terms of accuracy, there is not much difference between convolutional model and recurrent model. But the fact that CNN is faster than RNN training which is easy to understand and uncontroversial. The inevitable result is affected by two factors.

- Convolutions can be processed in parallel, and recurrent can only be processed sequentially. Having multiple machines trained parallel convolutional models simultaneously is certainly faster than training the serial recurrent model.
- The GPU chip can be used to speed up the training of the convolution model, and currently there is no hardware to speed up the RNN training.

The CNN-CNN based framework is a match between CNN and RNN in the field of machine translation and image captioning. In the recent years, CNN turns out to be a wide application given their effectiveness in computer vision and a lot of researches have been studied in the machine translation. In the same way, these improvements in convolutional model can be applied in the image captioning. Since the CNN-CNN framework in image captioning was first proposed in 2017, and there are many improvements using this framework in machine translation which can also be applied in image captioning. In the future study, more researches need to study in depth CNN-

CNN based attention mechanism and the combination of CNN and RNN in the decoder phase.

## 4 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 COCO dataset

| Method | Parameters | Time/Epoch |
|---|---|---|
| CNN-RNN [7] | 13M | 1529s |
| CNN-CNN [23] | 19M | 1585s |
| Reinforcement [21] | 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 Titan X 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.

## 5 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.

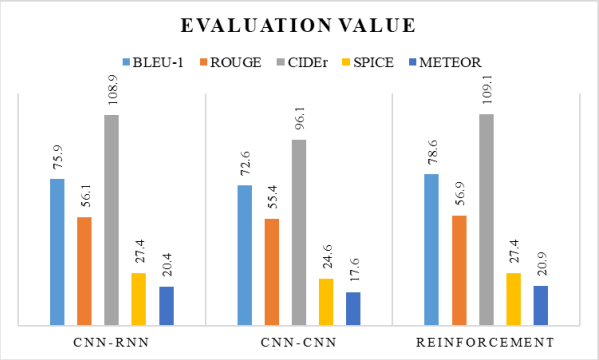


**Figure 5** Evaluation Index of three methods

In Fig.5, we have shown the best results of the above three methods for five evaluation metrics. We can see the both the CNN-RNN based and the Reinforcement based methods can get the better performance than the CNN-CNN based framework, which greatly improves the training speed without seriously affecting the accuracy. Besides, the reinforcement framework performances the best, since the objective function is more reasonable as we have introduced in Section 4.

# 6 Discussions

## 6.1 Benefits

If we are able to perform automatic image annotations, then this can have both practical and theoretical benefits. In the current social development process, the most important thing is the massive data that exists on the Internet. Most of these data are different from traditional data, and media data occupies a large proportion. They are often generated from Internet products such as social networks or news media. Apart from the fact that humans can directly process these media images, the useful information that the machine can currently collect from them is limited and it is difficult to assist humans in further work. Image captioning tasks, if they are accurate enough, can handle massive amounts of media data and generate human natural language descriptions that are more acceptable to humans. The machine will be able to better assist human beings to use these media data to do more things.

### *6.1.1 Intelligent monitoring*

Intelligent monitoring enables the machine to identify and determine the behaviour of people or vehicles in the captured scene and generate alarms under appropriate conditions to prompt the user to react to emergencies and prevent unnecessary accidents. For example, in channel monitoring, it collects the fairway operations and illegal activities, monitors the conditions of the fairway, and promptly discovers the conditions of the waterway operations, traffic conditions, illegal sand mining, and the use of navigation channels. Then report the situation to the command centre for scheduling and stop illegal activities in a timely manner. Image captioning can be applied to this aspect. Through the image captioning methods, the machine can understand the scenes it captures, so that it can respond to specific situations or notify users in a timely manner based on human settings.

### *6.1.2 Human-computer interaction*

With the advancements of science and technology and the need for the development of human life, robots have been used in more and more industries. Auto-pilot robots can intelligently avoid obstacles, change lanes and pedestrians based on the road conditions according to the surrounding driving environment they observe. In addition to safe and efficient driving, it is also possible to perform operations such as automatic parking. Liberating the driver's eyes and hands can greatly facilitate people's lives and reduce safety accidents. If the machine wants to do the work better, it must interact with humans better. The machine can tell humans what it sees, and humans then perform appropriate processing based on machine feedback. To complete these tasks, we need to rely on automatic generation of image descriptions.

### *6.1.3 Image and Video annotation*

When a user uploads a picture, the picture needs to be illustrated and annotated which can be easily found by the other users. The traditional method is to retrieve the most similar picture in the database for annotation, but this method often results in incorrectly annotated images. Besides, video has now become an indispensable part of people's lives. In order to enjoy movies better, many movies now require subtitles. Every year, there are a large number of videos produced worldwide. These videos are composed of tens of thousands of pictures. Therefore, image and video annotation are a heavy task. The automatic generation of the picture description can process all the video frames, and then automatically generate the corresponding text description according to the content of the video frame, which can greatly reduce the workload of the video worker and can complete the video annotation work efficiently and effectively. In addition, image and video annotation can also help visually impaired people to understand a large number of videos and pictures on the Internet.

The image description is generated automatically in the aspects of intelligent monitoring, human-computer interaction, image and video annotation. This is only part of the image captioning applications. In short, image captioning can indeed be applied in many aspects of people's lives, which can greatly improve labour efficiency and facilitate people's life, production and learning.

## 6.2 Major challenges

At present, the research of image captioning has gone through a long period of time and experienced several stages based on different technologies. Especially in recent

years, the application of neural network technology has opened up a new situation for image captioning research. Although the powerful data processing capability of neural network has a very outstanding performance in the study of image captioning generation, there are still some problems that have not been solved.

### *6.2.1 Richness of image semantics*

The current study can describe the image content to a certain extent, but it is not sensitive to the number of objects contained in the image. For example, the model often cannot accurately describe the objects with terms such as “two” or “group”. Besides, the selection of focus points in complex scenes are different. For people, it is easy to grasp the important content in the image and capture the information of interest. But for the machine, this will not be easy. The current image description automatic generation technology can describe pictures with simple scenes more comprehensively, but if the picture contains complex scenes and numerous object and object relationships, the machine often cannot grasp the important content in the image well. More attention will be paid to some minor information. This situation often affects the final result of the image description, sometimes even misinterpreting the original meaning of the image content.

### *6.2.2 Inconsistent objects during training and testing*

From the current study, during the training process, the input to the network at each time step is a real word vector or a mixture of real words and images, and the output of the network is the predicted word. However, in the test process, the network inputs at each time step is the output word vector in the vocabulary of the training dataset. The existing training process relies heavily on the selection of data sets. Once an given image contains novel objects, the approach taken is to select the closest object from the data set instead of the true object. In this way, there are inconsistencies in the training and the testing process when the new objects are created. Such inconsistencies may lead to the generation of cumulative error sampling, and even result in text descriptions that are completely inconsistent with the image content, resulting in incorrect description results.

### *6.2.3 Cross-language text description of images*

The existing image captioning method based on deep learning or machine learning requires a lot of marked training samples. In practical applications, it is required that a text description of a plurality of languages can be provided for the image to meet the needs of different native language users. At present, there are many training samples described in English and Chinese texts, but there are few mark-ups in other language text descriptions. If the textual description of each language in the image is carried out, manual marking will require a lot of manpower and time. Therefore, how to implement cross language text description of images is a key problem and a research difficulty in image captioning.

## 7 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.

## References

1. 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)
2. Girshick, Ross, et al. "Region-based Convolutional Networks for Accurate Object Detection and Segmentation." *IEEE Transactions on Pattern Analysis & Machine Intelligence* **38.1**:142-158. (2015)
3. Devlin, Jacob, et al. "Language Models for Image Captioning: The Quirks and What Works." *Computer Science* (2015)
4. Fang, H., et al. "From captions to visual concepts and back." *Computer Vision and Pattern Recognition IEEE*, 1473-1482. (2015)
5. Cho, Kyunghyun, et al. "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation." *Computer Science* (2014)
6. Hochreiter, Sepp, and J. Schmidhuber. "Long Short-TermMemory."*Neural Computation* **9.8**: 1735-1780. (1997)
7. Karpathy, Andrej, and F. F. Li. "Deep visual-semantic alignments for generating image descriptions." *Computer Vision and Pattern Recognition IEEE*, 3128-3137. (2015)
8. Sermanet, Pierre, et al. "OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks." *Eprint Arxiv* (2013)
9. Sundermeyer, M., et al. "Comparison of feedforward and recurrent neural network language models." *IEEE International Conference on Acoustics, Speech and Signal Processing IEEE*, 8430-8434. (2013)
10. Simonyan, Karen, and A. Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition." *Computer Science* (2014)
11. Szegedy, Christian, et al. "Going deeper with convolutions." *IEEE Conference on Computer Vision and Pattern Recognition IEEE*, 1-9. (2015)

12. He, Kaiming, et al. "Deep Residual Learning for Image Recognition." *IEEE Conference on Computer Vision and Pattern Recognition IEEE Computer Society*, 770-778. (2016)
13. Mao, Junhua, et al. "Explain Images with Multimodal Recurrent Neural Networks." *Computer Science* (2014)
14. Vinyals, Oriol, et al. "Show and tell: A neural image caption generator." *IEEE Conference on Computer Vision and Pattern Recognition IEEE Computer Society*, 3156-3164. (2015)
15. Xu, Kelvin, et al. "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention." *Computer Science* ,2048-2057. (2015)
16. Papineni, K. "BLEU: a method for automatic evaluation of MT." (2001)
17. Satanjeev, Banerjee. "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments." *ACL-2005*.228-231. (2005)
18. Flick, Carlos. "ROUGE: A Package for Automatic Evaluation of summaries." *The Workshop on Text Summarization Branches Out2004*:**10**. (2014)
19. Vedantam, Ramakrishna, C. L. Zitnick, and D. Parikh. "CIDEr: Consensus-based Image Description Evaluation." *Computer Science* ,4566-4575. (2014)
20. Anderson, Peter, et al. "SPICE: Semantic Propositional Image Caption Evaluation." *Adaptive Behavior* **11.4** 382-398. (2016)
21. Ranzato, Marc'Aurelio, et al. "Sequence Level Training with Recurrent Neural Networks." *Computer Science* (2015)
22. Kalchbrenner, Nal, E. Grefenstette, and P. Blunsom. "A Convolutional Neural Network for Modelling Sentences." *Eprint Arxiv* (2014)
23. Aneja, Jyoti, A. Deshpande, and A. Schwing. "Convolutional Image Captioning." (2017)
24. Gu, Jiuxiang, et al. "Stack-Captioning: Coarse-to-Fine Learning for Image Captioning." (2018)