Image Captioning based on Artificial Intelligence

Saurav Shaurya[1], Kunjal Prashant Shah[2], Gunda Umamaheshwar Gupta[3], Makineni Nagaswetha[4], Manish Kumar Govind[5], Naveed Hamid Merchant[6]

---***---

Abstract - With the development of deep learning , the aggregate of pc imaginative and prescient and natural language device has aroused great interest within the beyond few years. Image captioning is a representative of this filed, which makes the computer discover ways to use one or extra sentences to understand the seen content material of an picture. The vast description technology method of excessive degree image semantics calls for now not handiest popularity of the item and the scene, however the ability of reading the dominion, the attributes and the connection among the ones devices. Though photograph captioning is a complicated and tough project, a number of researchers have done sizeable enhancements. In this paper, we mainly describe 3 image captioning techniques the usage of the deep neural networks: CNN-RNN primarily based, CNN-CNN based totally and Reinforcement-based totally framework. Then we introduce the paintings of these 3 pinnacle techniques respectively, describe the assessment metrics and summarize the advantages and primary demanding situations.

Key Words: CNN, Computer Vision, RNN.

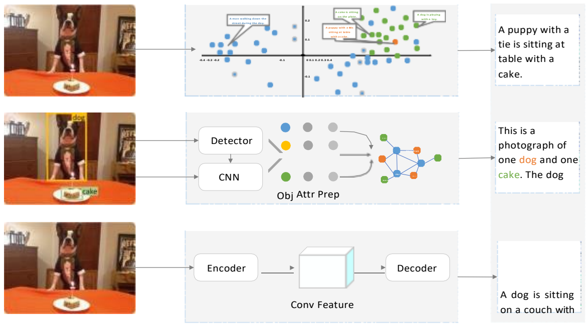
1.INTRODUCTION

In the past few years, pc imaginative and prescient in photograph processing area has made good sized development, like photograph kind [1] and item detection [2]. Benefiting from the advances of photo magnificence and object detection, it becomes feasible to mechanically generate one or extra sentences to apprehend the visible content of an photo, this is the hassle referred to as Image Captioning. Generating complete and Natural picture descriptions automatically has huge ability consequences, which incorporates titles attached to information images, descriptions associated with scientific photographs, textual content-based photograph retrieval, records accessed for blind clients, human-robot interplay. These packages in photo captioning have essential theoretical and sensible studies cost. Therefore, photo captioning is a more complicated but meaningful task within the age of artificial intelligence.

Given a latest picture, an picture captioning set of rules ought to output an define approximately this photo at a semantic diploma. For the image captioning undertaking, humans can without problems understand the picture content and explicit it inside the shape of Natural language sentences consistent with specific needs; however, for computers, it calls for the included use of photo processing, laptop imaginative and prescient, Natural language processing and distinctive major regions of research consequences. The task of picture captioning is to format a model that may fully use photograph statistics to generate greater human-like rich photograph descriptions. The significant description Era technique of excessive diploma photograph semantics calls for not simplest the information of items or scene reputation inside the photo, however additionally the capability to look at their states, recognize the relationship amongst them and generate a semantically and syntactically accurate sentence. It is presently doubtful how the mind is conscious an photo and organizes the seen records proper into a caption. Image captioning consists of a deep facts of the arena and which topics are salient elements of the whole.

1.1 Challenges

Despite such demanding situations, the problem has executed large improvements over the previous few years. Image captioning algorithms are typically divided into three training. The first magnificence, as demonstrated in , tackles this hassle the usage of the retrieval-primarily based techniques, which first retrieves the closest matching images, after which switch their descriptions because the captions of the query snap shots [3]. These strategies can produce grammatically correct sentences but cannot regulate the captions consistent with the ultra-modern photo. The 2nd magnificence , commonly uses template-primarily based strategies to generate descriptions with predefined syntactic rules and slit sentences into numerous elements [4]. These techniques first take gain of numerous classifiers to recognize the gadgets, in addition to their attributes and relationships in an picture, after which use a rigid sentence template to shape an entire sentence. Though it may generate a brand new sentence, those strategies both can not explicit the visible context efficiently or generate bendy and huge sentences


A man walking down the street during the day.
A cake is sitting on the plate.
A puppy with a tie sitting at table with a cake.
A dog is playing with a toy.
A puppy with a tie is sitting at table with a cake.
dog
cake
Detector
CNN
Obj Attr Prep
This is a photograph of one dog and one cake. The dog
Encoder
Conv Feature
Decoder
A dog is sitting on a couch with

With the extensive application of deep mastering, maximum current works fall into the 1/3 class known as neural community-primarily based techniques. Inspired through machine mastering's encoder-decoder structure [5], current years maximum image captioning techniques rent a

Convolutional Neural Network (CNN) because the encoder and a Recurrent Neural. Network (RNN) as the decoder, specifically Long Short-Term Memory (LSTM) [6] to generate captions [7], with the goal to maximize the likelihood of a sentence given the visual functions of an photo. Some methods are the use of CNN as the decoder and the reinforcement getting to know as the choice-making network.

According to those extraordinary encoding and decoding methods, on this paper, we divide the picture captioning strategies with neural networks into three classes: CNN-RNN based totally, CNN-CNN based totally and reinforcement-based framework for photograph captioning. In the subsequent component, we are able to speak approximately their most important thoughts.

2. CNN-RNN primarily based completely framework

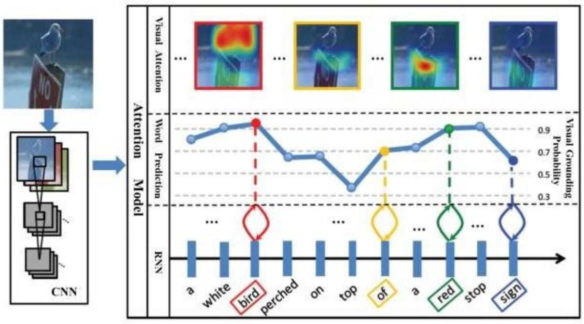
In human's eyes, an photo consists of various colorings to compose the wonderful scenes. But inside the view of laptop, maximum pics are painted with pixels in three channels. However, inside the neural community, wonderful modalities of statistics are all trending to create a vector and do the subsequent operations on those functions.

It has been convincingly proven that CNNs can produce a rich instance of the input picture by embedding it into a set-period vector, such that this example may be used for an expansion of imaginative and prescient responsibilities like item recognition, detection and segmentation [8]. Hence, picture captioning techniques based on encoder-decoder frameworks often use a CNN as an photo encoder. The RNN network obtains ancient records thru non-stop motion of the hidden layer, which has higher education talents and might perform better than mining deeper linguistic information together with semantics and syntax information implicit within the phrase collection [9]. For a dependency relationship between distinct region phrases in ancient statistics, a recurrent neural network may be results in easily represented inside the hidden layer usa. In image captioning assignment primarily based on encoder-decoder framework, the encoder detail is a CNN model for extracting photograph capabilities. It can use fashions which includes AlexNet [1], VGG [10], GoogleNet [11] and ResNet [12]. In the decoder factor, the framework enters the phrase vector expression into the RNN model. For every phrase, it is first represented by way of a one-hot vector, after which via the word embedding model, it will become the equal size due to the fact the picture characteristic. The photograph captioning hassle can be described in the shape of a binary (I, S), wherein

I represents a graph and S is a chain of target words, S = S1, S2 ⋯ and Si is a phrase from the facts set extraction. The purpose of schooling is to maximize the hazard estimation of the target description $p(S$goal of the generated statement and the target statement

matching extra intently. Mao et al. [13] proposed a multimodal Recurrent Neural Network(m-RNN) version that creatively combines the CNN and RNN version to clear up the photograph captioning hassle. Because of the gradient disappearance and the constrained reminiscence problem of normal RNN, the LSTM version is a unique form of shape of the RNN model that could solve the above issues. It gives three manipulate devices (cellular), that are the enter, output and forgot gates. As the facts enters the model, the information may be judged via way of the cells. Information that meets the rules might be left, and nonconforming records can be forgotten. In this precept, the lengthy collection dependency trouble within the neural community can be solved. Vinyals et al. [14] proposed the NIC (Neural Image Caption) model that takes an photograph as enter in the encoder detail and generates the corresponding descriptions with LSTM networks within the decoder part. The version solves the problem of vectorization of Natural language sentences thoroughly. It is of super importance to use computers dealing with herbal language, which makes the processing of computers no longer remains at the smooth degree of matching, but in addition to the extent of semantic information.

Inspired via the neural community-based device translation framework, the attention mechanism inside the discipline of laptop vision is proposed to sell the alignment between phrases and photo blocks. Thereby, in the gadget of sentence technology, the "attention" transfer way of simulating human imaginative and prescient may be collectively promoted with the generation technique of the word sequence, in order that the generated sentence is greater consistent with the human beings's expression dependancy. Instead of encoding the entire image as a static vector, the eye mechanism gives the complete and spatial information much like the image to the extraction of the image features, resulting in a richer assertion description. At this time, the photo features are considered because the dynamic function vectors combined with the weights information. The first interest mechanism changed into proposed in [15], it proposed the "smooth attention" this means that that to pick out regions primarily based totally on special weights and the "hard interest" which performs interest on a selected visible concept. The experimental outcomes obtained with the useful resource of the use of attention-based deep neural networks have carried out exceptional consequences. Using interest mechanism makes the model generate every word steady with the corresponding location of an photograph as is proven.



However, it also suffers from predominant drawbacks For the image captioning venture, which encourage further tremendous studies. The first is that the metrics used for trying out and loss for education are one-of-a-type. We use skip- entropy as loss, however metrics are non-differentiable and can not be straight away used as schooling loss. And log opportunity may be visible as giving the same weight to each word, but in fact humans compare considered one of a type words with selective weights. This discrepancy is referred to as "loss-evaluation mismatch" trouble [21]. The 2d is that after schooling, the enter of each time step comes from the real caption and when generated, each phrase generated is based totally at the formerly generated phrase; Once a phrase isn't always generated nicely, it may get some distance far from the floor truth. This discrepancy is called "publicity bias" hassle [21].

3. CNN-CNN based totally absolutely framework

Although models like LSTM networks have reminiscence cells that can memorize the prolonged information records of the collection technology system higher than RNN, it is nonetheless updated at whenever, which render the lengthy-time period reminiscence as an opportunity hard. Inspired with the aid of the paintings of system gaining knowledge of, recent works have proven blessings of CNN at the photo captioning paintings. Using the CNN in NLP for text generation has been proved very effective [22]. In the sphere of neural system translation, it has proved that the CNN convolution model is used to update the RNN recurrent version, which not best exceeds the accuracy of the cycle model, however moreover will increase the education pace by using a element of nine. Most picture captioning works are stimulated with the aid of the gadget translation, due to the fact the translation art work is in the sequence to series structure and in the photograph captioning project, an photograph is considered as a sentence in a source language. To the excellent of our information, the first convolutional network for the textual content technology approach in image captioning is the paintings completed through Aneja et al. [23] and we name this CNN-CNN based framework.

This framework includes three number one components similar to the RNN technique. The first and the last additives are phrase embeddings in each instances. However, on the identical time as the centre element includes LSTM or GRU (Gated Reccurent Unit) gadgets inside the RNN case, masked convolutions are employed in the CNN-based definitely technique. This issue, unlike the RNN, is feed-beforehand with none recurrent feature. Aneja et al. [23] has installed the CNN-CNN framework has a faster schooling time according to amount of parameters but the loss is higher for CNN than RNN. The cause of the CNN version's accuracy is that CNN are being penalized for generating an awful lot much less-peaky word possibility distributions. However, plenty less peaky distributions aren't necessarily awful, wherein more than one phrase predictions are feasible for predicting diverse captions

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.

Actually, the layered abstraction of convolution and the triple gate of recurrence play the not unusual role. Although the technique are one among a kind, the reason is to disregard minor content material and spotlight the vital content material cloth. Therefore, in phrases of accuracy, there isn't always a outstanding deal distinction between convolutional version and recurrent model. But the truth that CNN is faster than RNN schooling which is straightforward to understand and uncontroversial. The inevitable result is suffering from two elements.

 Convolutions can be processed in parallel, and recurrent can best be processed sequentially. Having more than one machines educated parallel convolutional models simultaneously is genuinely quicker than education the serial recurrent model.

 The GPU chip may be used to hurry up the education of the convolution model, and currently there is no hardware to hurry up the RNN education.

The CNN-CNN based totally framework is a match between CNN and RNN in the field of machine translation and photograph captioning. In the cutting-edge years, CNN appears to be a massive software given their effectiveness in laptop vision and severa researches had been studied in the system translation. In the equal manner, those improvements in convolutional version may be completed in the image captioning. Since the CNN-CNN framework in image captioning became first proposed in 2017, and there are numerous upgrades the use of this framework in tool translation which also can be applied in image captioning. In the future have a look at, more researches need to look at giant CNN-based totally completely interest mechanism and the mixture of CNN and RNN inside the decoder segment.

4. Reinforcement primarily based framework

Reinforcement mastering has been extensively utilized in gaming, control idea, and so on. The issues on pinnacle of factors or gaming have concrete desires to optimize by using the usage of nature, while defining the precise optimization intention is nontrivial for image captioning.

When making use of the reinforcement studying into photo captioning, the generative model (RNN) may be regarded as an agent, which interacts with the outside surroundings (the phrases and the context vector because the enter at each time step) . The parameters of this agent define a coverage, whose execution results within the agent choosing an motion. In the collection era putting, an movement refers to predicting the subsequent word in the collection at every time step. After taking an movement the agent updates its internal nation (the hidden devices of RNN) . Once the agent has reached the stop of a sequence, it observes a praise. In this form of framework, the RNN decoder acts like a stochastic insurance, where selecting an movement corresponds to producing the subsequent phrase. During schooling PG technique chooses actions constant with the modern-day policy and most effective take a look at a reward on the cease of the gathering (or after most collection period) , via evaluating the collection of movements from the present day-day policy in competition to the most effective movement series. The purpose of training is to discover the parameters of the agent that maximize the predicted reward.

The idea of using PG (policy gradient) to optimize non differentiable dreams for photograph captioning was first proposed inside the MIXER paper [21], via treating the rating of a candidate sentence as analogous to a praise signal in a reinforcement analyzing putting. In the MIXER technique, for the purpose that hassle setting of textual content technology has a totally massive motion space which makes the hassle be hard to take a look at with an preliminary random policy, it takes moves of education the RNN with the cross-entropy loss for numerous epochs using the ground fact sequences which makes the version can consciousness on a notable a part of the quest space. This is a contemporary form of training that mixes collectively the MLE (maximum hazard estimation) and the reinforcement goal. This reinforcement getting to know version is pushed with the useful resource of visible semantic embedding, which plays properly for the duration of considered one of a type assessment metrics without re-training. Visual- semantic embedding, which affords a degree of similarity between photos and sentences, can degree similarities among photos and sentences, the correctness of generated captions and serve an inexpensive international purpose to optimize for photo captioning in reinforcement gaining knowledge of. Instead of learning the sequential loop model to greedily discover the subsequent accurate phrase, the selection-making community uses the "coverage network" and the "fee community" to at the same time determine the following quality phrase for whenever step. The policy network presents the confidence of predicting the subsequent phrase consistent with modern-day country. The price network evaluates the reward price of all possible extensions of the modern 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 examine the training parameters and training time (in seconds) for RNN, CNN and Reinforcement Framework. The timings are obtained on Nvidia Titan X GPU. We teach a CNN faster in keeping with parameter than the RNN and Reinforcement framework. But as for the accuracy and the range, the general overall performance of CNN is worse than the opposite models, this is illustrated in the following section.

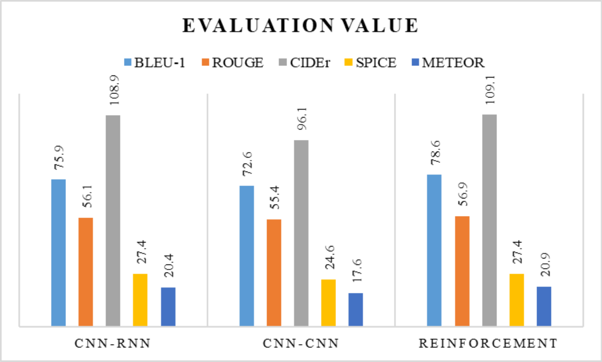
5. Evaluation metrics

The cutting-edge study on the whole uses the degree of matching between the caption sentence and the reference sentence to evaAluate the professionals and cons of the generation outcomes. The typically used strategies embody BLEU [16], METEOR [17], ROUGE [18], CIDEr [19], and SPICE [20] those 5

dimension signs. Among them, BLEU and METEOR are derived from system translation, ROUGE is derived from text abstraction, and CIDEr and SPICE are particular signs based totally on photo captioning.

 BLEU is extensively used within the evaluation of photo annotation consequences, this is based at the n-gram precision. The precept of the BLEU degree is to calculate the distance some of the evaluated and the reference sentences. BLEU technique has a tendency to provide the better rating whilst the caption is closest to the duration of the reference assertion.

 ROUGE is an automated assessment elegant Discussions.



6. Benefits

If we're able to carry out automatic image annotations, then this will have each practical and theoretical advantages. In the modern social improvement approach, the maximum crucial issue is the huge data that exists at the Internet. Most of these facts are one-of-a-type from traditional facts, and media information occupies a big proportion. They are often generated from Internet merchandise collectively with social networks or facts media. Apart from the truth that humans can at once approach the ones media images, the beneficial statistics that the device can currently collect from them is constrained and it's miles hard to assist human beings in in addition paintings. Image captioning responsibilities, if they're correct enough, can address massive quantities of media information and generate human natural language descriptions that are extra best to humans. The gadget might be capable of better assist human beings to use the ones media records to do more subjects.

6.1 Intelligent tracking

Intelligent monitoring lets in the system to perceive and decide the behaviour of human beings or cars within the captured scene and generate alarms below appropriate situations to spark off the consumer to react to emergencies and prevent useless injuries. For example, in channel tracking, it collects the fairway operations and unlawful activities, video display units the situations of the inexperienced, and at once discovers the conditions of the waterway operations, site visitors situations, unlawful sand mining, and the usage of navigation channels. Then document the situation to the command centre for scheduling and prevent unlawful sports in a well timed way. Image captioning can be carried out to this aspect. Through the photo captioning strategies, the device can apprehend the scenes it captures, in order that it could respond to specific conditions or notify customers in a well timed manner based on human settings.

6.2 Human-pc interaction

With the upgrades of technological information and generation and the want for the improvement of human life, robots were used in increasingly industries. Auto-pilot robots can intelligently keep away from obstacles, change lanes and pedestrians primarily based on the road situations in line with the encompassing riding environment they have a look at. In addition to safe and efficient using, it is also viable to perform operations along side automated parking. Liberating the motive pressure's eyes and palms can drastically facilitate humans's lives and reduce protection injuries. If the system wants to do the work better, it need to engage with humans higher. The gadget can tell people what it sees, and human beings then carry out appropriate processing primarily based on device remarks. To entire those obligations, we need to depend upon automatic generation of photo descriptions.

6.3 Image and Video annotation

When a consumer uploads a photo, the picture wishes to be illustrated and annotated which can be without problem observed with the aid of the other users. The traditional technique is to retrieve the most similar photograph in the database for annotation, however this technique regularly results in incorrectly annotated pix. Besides, video has now end up an vital part of human beings's lives. In order to enjoy films higher, many films now require subtitles. Every yr, there are a massive quantity of motion photographs produced worldwide. These films are composed of tens of heaps of images. Therefore, photo and video annotation are a heavy project. The automatic era of the picture description can method all of the video frames, and then robotically generate the corresponding text description in keeping with the content of the video frame, which could drastically reduce the workload of the video worker and can whole the video annotation artwork efficiently and efficaciously. In addition, image and video annotation also can help visually impaired people to recognize a huge kind of videos and pictures on the Internet.

The picture description is generated automatically in the elements of wise monitoring, human-pc interaction, photograph and video annotation. This is handiest a part of the photograph captioning packages. In quick, photo captioning can certainly be executed in plenty of elements of humans's lives, that could substantially decorate labour performance and facilitate human beings's life, manufacturing and reading.

6.4 Major demanding conditions

At gift, the research of photograph captioning has long past through an extended time frame and professional numerous levels based on specific technology. Especially in modern years, the utility of neural network era has opened up a modern day scenario for picture captioning research. Although the powerful information processing functionality of neural network has a completely exceptional overall performance in the examine of photo captioning generation, there are however some troubles which have not been solved.

6.4.1 Richness of photo semantics

The present day have a study can describe the image content material cloth to a effective quantity, however it isn't sensitive to the variety of gadgets contained within the image. For instance, the model frequently can't as it must be describe the devices with phrases together with "two" or "enterprise". Besides, the choice of focus elements in complicated scenes are unique. For human beings, it is simple to comprehend the vital content fabric inside the photo and capture the information of hobby. But for the gadget, this could no longer be easy. The present day image description computerized era technology can describe images with easy scenes more comprehensively, but if the photo contains complex scenes and numerous item and item relationships, the device often cannot draw close the critical content fabric within the image well. More interest can be paid to a few minor information. This scenario often influences the final result of the photo description, from time to time even misinterpreting the actual because of this of the photograph content material material.

6.4.2 Inconsistent items sooner or later of schooling and sorting out

From the cutting-edge observe, at some level inside the training process, the input to the network at on every occasion step is a real phrase vector or a mixture of real phrases and images, and the output of the community is the expected phrase. However, inside the test machine, the network inputs at on every occasion step is the output word vector within the vocabulary of the education dataset. The current education device is primarily based closely on the selection of facts units. Once an given image includes novel devices, the approach taken is to select the nearest object from the records set rather than the genuine object. In this way, there are inconsistencies inside the education and the trying out technique while the new devices are created. Such inconsistencies may additionally additionally purpose the generation of cumulative mistakes sampling, or maybe bring about text descriptions which can be simply inconsistent with the photograph content fabric, ensuing in incorrect description effects.

6.4.3 Cross-language text description of photographs

The present day image captioning approach primarily based on deep reading or machine studying calls for a whole lot of marked schooling samples. In sensible programs, it's far required that a text description of a plurality of languages can be supplied for the photograph to meet the needs of diverse local language customers. At present, there are various schooling samples defined in English and Chinese texts, however there are few mark-america of americain one-of-a-kind language textual content descriptions. If the textual description of each language inside the photo is carried out, guide marking might require hundreds of manpower and time. Therefore, the way to put in force skip language textual content description of pix is a key problem and a research difficulty in photo captioning.

7. Conclusion

Image captioning has made enormous advances in latest years. Recent work primarily based on deep analyzing techniques has caused a breakthrough within the accuracy of photo captioning. The text description of the image can decorate the content material cloth-based totally image retrieval efficiency, the growing software program scope of visual know-how within the fields of medication, security, military and distinctive fields, which has a big application prospect. At the equal time, the theoretical framework and studies strategies of photograph captioning can promote the improvement of the theory and alertness of photograph annotation and visible question answering (VQA), go media retrieval, video captioning and video conversation, which has critical academic and practical application value.Designed to evaluate textual content summarization algorithms. There are 3 evaluation criteria, ROUGE-N, ROUGE-L, and ROUGE-S. ROUGE-N is commonly basedon the given sentence to be evaluated, which calculates a simple n-tuple recollect for all reference statements: ROUGE-L is primarily based totally on the biggest not unusual collection (LCS) calculating the don't forget. ROUGE-S calculates undergo in thoughts based mostly on co- occurrence information of bypass-bigram among reference textual content description and prediction textual content description.

 CIDEr is the unique method this is furnished for the picture captioning work. It measures consensus in image captioning via acting a term frequency inverse file frequency (tf-idf) for each n-gram. Studies have verified that the suit among CIDEr and human consensus is better than distinctive assessment criteria.

 METEOR is based totally mostly on the harmonic suggest of unigram precision and recall, but the weight of the undergo in thoughts is better than the accuracy. It is extraordinarily relevant to human judgment and differs from the BLEU in that it isn't always best within the entire set, but moreover in the sentence and segmentation degrees, and it has a excessive correlation with human judgment.

 SPICE evaluates the high-quality of photograph captions with the resource of changing the generated description sentences and reference sentences into graph-based totally semantic representations, particularly "scene graphs". The scene graphs extract lexical and syntactic records in

Natural language and explicitly represents the gadgets, attributes, and relationships contained inside the photo

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)