JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

EXPLORING DEEP LEARNING PARADIGMS FOR IMAGE CAPTIONING

[1]Chandan Shukla, [2]Aditya Agarwal, [3]Harshit Kumar Rai, [4]Rishabh Joshi, [5]Prabhat Singh

[1]Student, [2]Student, [3]Student, [4]Student, [5]Assistant Professor
[1]Dept. of Computer Science & Engineering (Data Science),
ABES Engineering College, Ghaziabad, India

Abstract: There has been a significant increase in interest in bringing together computer vision and natural language processing since the announcement of deep learning. Picture captioning serves as a representation for this field; it uses one or more sentences to teach a computer how to understand the visual information in an image. In order to provide meaningful descriptions for high-level image semantics, one must also be able to analyze the state, attributes, and relationships between these objects. Though picture captioning remains a difficult and complex task, several researchers have made significant progress. Three deep neural network-based image captioning techniques—RNN, CNN-CNN, and reinforcement learning frameworks—are the main topics of discussion in this study. Next, we outline the key benefits and challenges, go over the assessment criteria in brief, and offer sample work for each of the top three techniques.

Keywords – Image Captioning, Deep Learning, CNN, RNN, LSTM, Encoder-decoder architectures, Training Methodologies, Flickr8k Dataset, Performance Benchmarks

I. Introduction

Recent advances in computer vision and image processing may be seen in the areas of picture categorization [1] and object identification [2]. The problem of picture captioning, which is the automatic generation of one or more lines that talk about the main content of an image, has been helped by advances in object identification and image classification. The ability to automatically generate comprehensive and natural picture descriptions might have a wide range of effects, such as text-based image retrieval, medical image descriptions, news image titles, information accessible to blind users, and human-robot interaction. Users who are blind, as well as human-robot interaction. These captioning-related apps are truly useful. for both theoretical and practical research. Hence, picture captioning is a crucial task in the era of artificial intelligence.

When a new image is provided, an image captioning algorithm should give out a semantic explanation of it. For example, the input image in Fig. 1 is composed of humans, surfing planks, and surging tides. A statement at the footer describes the substance of a picture. The scene, the activity, and the objects that are seen in the picture are all described in one statement.

The task of captioning images for a computing machine requires the processing of pictures, computing natural language, computer vision, etc. Contrastingly, people can understand items in a picture, express as a decision with the help of NLP. High level picture semantics requires the ability to assess the states of the objects or scenes in the image, understand their relationships, and generate a phrase that is correct both syntactically and semantically in order to offer a meaningful description. The process by which the brain understands is yet unknown.

Notwithstanding these challenges, the problem has advanced significantly in recent years. Three different kinds of picture captioning algorithms are typically used. The initial group, seen in Figure 2.(a), addresses this problem by employing retrieval-based techniques, which get the pictures that most closely match the query images first, then append their captions to the query images [3].

a couple of surfers standing on their boards.

fig 1: an example of description about a picture

There are many things that can give text, yet they cannot make the captions match the new image. Fig. 2's second category. Template-based techniques are frequently used in (b), which creates descriptions using preset grammatical rules and breaks sentences up into many parts [4]. All processes start by identification of items in an image, together with attributes along with respective connections, using a variety of classifiers.

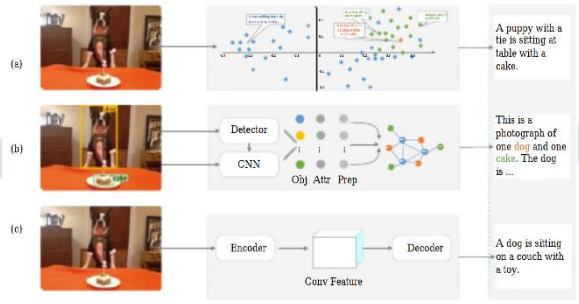


fig 2: 3 distinct groups for picture interpretation.

Because deep learning (c) is being used widely, the majority of current efforts come in the category of AI-based techniques in Fig2. To maximize the probability of a phrase given the visible characteristics of a picture, most image captioning techniques today use a RNN for decoder, also a CNN for encoder. In particular, Short-Term Memory [6] serves like processor for caption generator [7]. Encoder-decoder architecture of machine learning serves as an inspiration for this method [5]. Reinforcement learning is sometimes utilized as the decision-making network, and CNN is occasionally employed as the decoder.

Based on the various encrypting and decrypting ways, we split up the neural network-based picture processing approaches into these multiple ways. They will follow the address with their key points.

II. Literature Review

The transition from rule-based to neural network-based models in this field signified a paradigm change from earlier techniques.

Earlier picture captioning systems used preset templates and were rule-based. In this context, "Describing Visual Scenes" by K. Barnard et al. (1997) is a noteworthy article that examined rule-based caption generation techniques. These methods, however, were not able to handle a wide range of pictures or capture intricate connections.

Improvements in picture captioning were sparked by the introduction of deep learning. The work of A. Krizhevsky et al. in "ImageNet Classification with Deep CNN" (2012) was noteworthy in this shift. The efficacy which CNN has in picture categorization was established in this article, opening the door for its application in picture captioning.

Recurrent Neural Networks (RNNs) for caption generation and CNNs for image feature extraction became the usual method. In "Show and Tell: A Neural Image Caption Generator" (2015), Vinyals et al. presented an end-to-end trainable model and introduced the idea of utilizing neural networks for both caption production and picture encoding.

III. Framework based on CNN-RNN

Inside the eyes of people, a profile is made in several tints that depict different scenes. Nonetheless, when seen on a computer, the majority of pictures are produced using pixels in three channels. On the other hand, different neural network data modalities are all working towards the same goal, which is to create a vector and then operate with characteristics.

The rich portrayal is utilized in various visual works, including body identification, segmentation, and acceptance [8]. Consequently, CNN is often employed as a picture encoder in encoding-decoding framework-themed image caption generating systems. Recurrent Neural network regularly rotates its concealed layer to obtain previous information because this provides greater training potential for performing exclusively and thorough knowledge [9].

In a few hidden layer states, RNN may be readily explained for a dependent connection between many location terms in past data. An encoder-decoder system for captioning pictures uses a CNN model for feature extraction as the encoder. Models like Res Net [12], Google Net [11], VGG [10], and Alex Net [1] can be employed by it. Inside the decrypting step, the structure feeds

declaration of representation. Thanks to the phrase embedding approach, every term is shown by a single point becomes unchanged as the picture characteristics. An expression for a photo indication is a dual problem.

Mao et al. [13] developed a multi-channel m-RNN type which cleverly fuses the Convolutional and Recurrent Neural Network models in tackling of photo. The gradient disappearance and restricted memory problems of the conventional RNN may be solved by the LSTM representation, a unique kind of structure of the Recurrent Neural model. There are three more control units (cells) added: forget gates, input, and output. The model's cells will assess the data as it comes in. Information that conforms with the criteria will be kept, while nonconforming content will be deleted. This technique may be used to fix the neural network's protracted sequence dependency problem. Vinyals et al. introduced NIC (Neural Image Caption) paradigm. To generate the matching descriptions, a picture is fed into encrypter, and Long Short Term memories are used in decrypter. The problem of vectorizing phrases in natural language is efficiently solved by the model. NLP is a critical area in computer science because it takes computer science beyond simple matching and towards semantic comprehension.

In computer vision, the attention approach is introduced to promote word alignment and picture block alignment. It draws inspiration from the machine translation system using NN. The process of generating the word sequence can mutually stimulate the "focus" shift method for replicating people eyesight, so that the created phrase is more in accordance with people's expressive patterns. Rather than collecting the entire image as a fixed line, the recognition method integrates the image's entirety of location data to the drawing out of picture characteristics, resulting in a deeper line definition. As of right now, weights data and dynamic feature vectors are used to describe the image features.

The initial focus system was put forth in [15]. This suggested "soft attention," this entails selecting areas depending on various measurements, and "hard attention," involving focusing on a specific visualizing idea. Deep neural networks with an attentional focus have produced impressive results from the experiment. As shown in Figure 3, the framework creates every word based on comparable area of an image using an attention process.

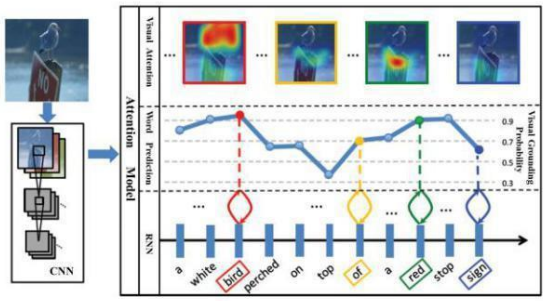


fig 3: demonstration of attention mechanism

Complexities are used in Convolutional NNmethods. These elements are data-forwarded and does not have recurrent capabilities as the RNN does. Aneja et al. [23] report that the Convolutional NN-Convolutional NN architecture has a lower practice work per parameter, even though CNN has a greater loss than Recurrent NN. Whereas, Fig. 4 shows, less peaky distributions are not always a negative thing because they enable different word predictions to be made for predicting a range of captions.

For the picture captioning assignment, it also has two significant shortcomings that need for more investigation. Because metrics lack differentiability, we need to employ cross-entropy as the training loss instead of directly utilizing metrics. Moreover, although log likelihood could seem that every word has the same weight, individuals really assign different weights to different words based on their individual evaluations. This disparity is also called the "loss-evaluation mismatch" issue [21]. A word is not made correctly, every word that is formed during training is based on a previous word that was generated. This means that work given for every point is coming out of genuine inscription.

IV. Framework based on CNN-CNN

Long-term memory known a bit difficult because, although models such as Long Short Term Memory networks include spaces that store rich historical data steps building action more effectively than Recurrent NN. Using Convolutional NN for word creation has proved to be quite successful [22].

In AI translation, studies have demonstrated that convolutional neural models can effectively replace recurrent neural models, offering improved accuracy and accelerated training by a significant factor. All efforts to create captions for images are influenced by computer interpretation since most of the process of translating is sequential framework, where each picture is treated as a phrase in the original speech. We cite this framework as Convolutional-Convolutional neural network based because, to our current understanding, Aneja et al.'s research [23] stands out as the leading convolutional network for phrase synthesis in photo commenting tasks.

Three key elements of this system resemble the RNN methodology. In both scenarios, the initial and final components consist of word embeddings. However, in the RNN setup, masked LSTM or GRU (Gated Recurrent Unit) units are situated in the central component.

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.

fig 4: produced explanation for different models

The triple gate mechanism of recurrence and the layered abstraction of convolution serve analogous purposes. Both aim to disregard irrelevant information and highlight pertinent content, irrespective of their respective methodologies. Consequently, there exists minimal distinction between the accuracy of recurrent and convolutional models. However, it is clear and undisputed that learning to operate convolutional is quicker than recurrent model. There are two things that influence the unavoidable outcome.

Recurrence necessitates sequential handling, whereas convolutions allow for parallel processing. Training parallel convolutional models across multiple machines is unquestionably faster compared to training a serial recurrent model.

The convolution model's training can be accelerated using the GPU chip, although there isn't any hardware available right now to accelerate Recurrent Neural Network training.

In areas of machine translation and image captioning, CNN and RNN are matched by the CNN-CNN based architecture. Given its success in computer vision and the extensive research that has been done on machine translation, CNN has recently emerged as a popular application. These enhancements to the convolutional model can also be used for picture commenting. Since the introduction of the CNN-CNN framework for image captioning in 2017, machine translation has undergone numerous advancements that are applicable to enhancing image captioning techniques. Further research is required in the future to comprehensively investigate Convolutional-convolutional focused attention systems and the incorporation of CNN and RNN architectures during the decoding phase.

V. Reinforcement based framework.

Reinforcement learning has been extensively employed in gaming. However, establishing an appropriate optimization objective for picture captioning is not as straightforward as it is for control or gaming problems, which inherently possess tangible targets for improvement.

In annotating photographs using reinforcement learning, the generative algorithm (a neural network, RN) acts as an agent interfacing with the external world. It continuously receives inputs from text and contextual variables at each step of the process. This agent's variables specify a rule, which when performed leads the agent to take certain steps. Anticipating the subsequent phrase at each stage constitutes a step within the sequence generation process. Subsequent to an action, the agent, represented by the hidden unit of the RNN, adjusts its internal state. Upon completion of an order, the agent detects an incentive. On this type of situation, the recurrent neural processor operates as a probabilistic strategy, where selecting a single choice result in the creation of the following text. The Policy Gradient (PG) technique selects actions during training that align with the current strategy and only receives rewards at the end of the process (beyond the overall performance metric). This is achieved by comparing the sequence of actions taken by the current policy to the optimal action sequence. The training objective is to identify policy parameters that maximize the expected reward.

It was MIXER research [21] that first proposed implementing policy gradient to maximize nondifferentiable targets for picture captioning. This study compared an applicant phrase's value to a payout message in an interactive reinforcement learning context. Because text generation problem setting contains fairly big activity area and is challenging to learn with an initial random policy, the MIXER method uses actions to train RNN using cross-entropy degradation for numerous epochs utilizing genuine sequences. This allows the algorithms to concentrate over significant percentage of the domain of search. This novel approach to training combines the estimation of maximum likelihood with reinforced targets.

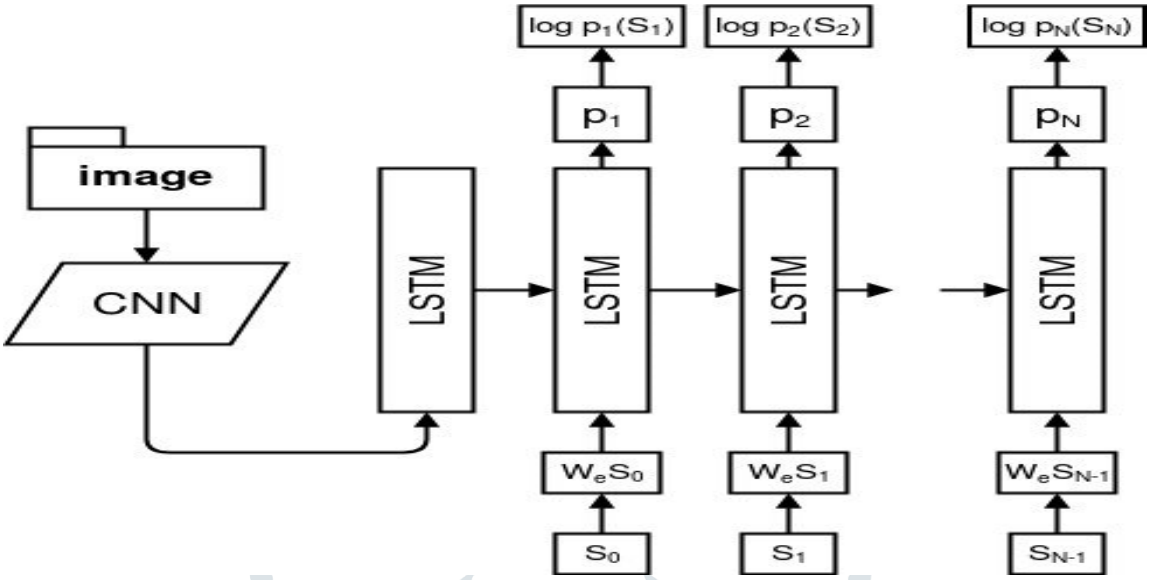
The reinforcement-based form of instruction taking model performs better than conventional evaluation measures without requiring retraining thanks to its visual semantic embedding. In addition to evaluating the quality of produced captions, visual-semantic embedding—which gauges the similarity between words and images—makes a solid starting point for picture captioning optimization in reinforcement learning.

Instead of relying solely on a systematic looping approach to determine the next correct phrase in a greedy manner, choice-making networks utilize both a "policy network" and a "value network" to collaboratively select the most suitable phrase at each

stage. As of right now, the policy network provides the assurance needed to predict the next word. The value network assesses the potential reward value associated with all potential outcomes stemming from the present state.

Timings were acquired using the Nvidia MX250 GPU. It's noted that training a CNN per parameter tends to be faster compared to training both RNN and Reinforcement frameworks. Nevertheless, the subsequent section illustrates that CNNs exhibit lower performance in both variety and accuracy compared to other models.

VI. Methodology



VII. Evaluation metrics

The present research primarily utilizes the level of similarity among reference phrases and the caption phrase to calculate benefits and drawbacks of producing outcomes. The top five commonly used measurement indicators include Bilingual Evaluation Understudy [16], Metric for Evaluation of Translation with Explicit Ordering, Recall-Oriented Understudy for Gisting Evaluation, Consensus-based Image Description Evaluation, and Semantic Propositional Image Caption Evaluation. This includes the machine translation-derived BLEU and METEOR, the text abstraction-derived ROUGE, and the image captioning-based CIDEr and SPICE.

BLEU is widely used for n-gram precision-based image annotation outcomes assessment. The BLEU metric calculates the distance between the evaluated and reference phrases, often yielding a higher score when the caption closely matches the length of the reference sentence.

ROUGE, an automated assessment metric, is specifically designed to evaluate text summarization systems. It encompasses three evaluation criteria: Recall-Oriented Understudy for Gisting Evaluation - N-gram, Recall-Oriented Understudy for Gisting Evaluation - Longest Common Subsequence, and Recall-Oriented Understudy for Gisting Evaluation - Skip-bigram.

The fundamental computation in ROUGE-N evaluates n-tuple recalls for each citation statement. Recall-Oriented Understudy for Gisting Evaluation - Longest Common Subsequence employs LCS to calculate recall, while Recall-Oriented Understudy for Gisting Evaluation - Skip-bigram computes recall by considering the combination of skip-bigrams between the reference summary and the predicted text.

METEOR relies by the harmonic mean of monogram recollection and accuracy, with recall being assigned greater importance than precision. It is distinct from the BLEU in that it has a significant link with human assessment and is present not just within the overall collection, but also the phrase as well as segment stages. This additionally involves a great deal to do with human judgement.

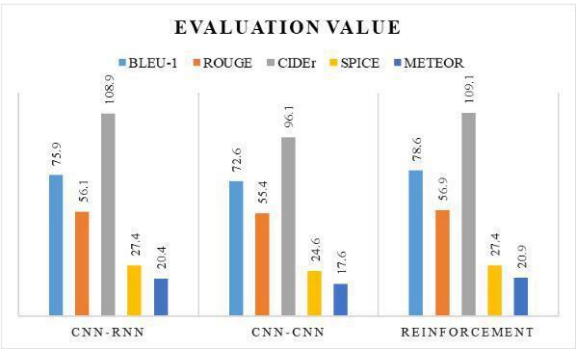


fig 5: three methods should be evaluated based on their respective performance using evaluation indices.

We have demonstrated the most favorable outcomes among all aforementioned techniques for five assessment metrics in Fig. 5. It is evident that, despite retaining a high degree of accuracy, the framework based on the CNN-CNN has outperformed in terms of performance by both the Reinforcement based and CNNRNN based methods. Furthermore, the goal function is more logical for the reinforcement structure, as we covered in Section 4, which explains why it performs the best.

VIII. Discussions

8.1 Benefits

Both theoretically and practically, the ability to automatically annotate photographs might be helpful. The most important component of the contemporary societal growing phase is large quantity of information accessible on the Internet. A significant portion of these data are media data, and the bulk are not standard data. Online businesses like social networks and news sites usually develop them. Apart from the fact that these media pictures may be directly interpreted by people, the extraction of valuable details by machines at present faces restrictions, posing challenges for future assistance to humans in their work.

8.1.1 Smart Surveillance

With smart supervision, gadgets can identify and evaluate movements of individuals or cars in recorded site. In some situations, it can also trigger alerts that will notify the user to take emergency action and prevent preventable collisions. For example, channel monitoring collects information on illegal activities and fairway operations, monitors the condition of the fairway, rapidly ascertains traffic patterns, illicit the extraction of sand and utilization of shipping routes. Subsequently, cease any unlawful behavior right away and notify the problem to the command center so that a time may be set for it. Image captioning can be used to alleviate this issue.

8.1.2 Interaction between humans and computers

These days, a growing number of sectors utilize robots because of scientific and technological advancements as well as the need to sustain human existence. Autonomous robots have the ability to sense road conditions and make intelligent decisions, such as avoiding obstructions, changing lanes, and avoiding pedestrians. It is possible to drive safely and efficiently and to automate processes like parking. If the driver can keep their hands and eyes free, it can significantly enhance people's life and reduce safety incidents. Enhancing human-computer interaction is necessary to maximize the computer's performance. When the machine reports what it has seen, humans may utilize its input to direct their own processing.

8.1.3 Annotation of Images and Videos

A picture that a user contributes has to be categorized and annotated in order for other users to locate it easily. The traditional method involves searching the database for the closest comparable image and retrieving it for explanation; however, this approach frequently yields inaccurately labeled images. Furthermore, video has evolved a necessity in people's lives. Subtitles are now necessary for many films in order to fully enjoy them. Each year, a significant amount of clips are made globally. The films aforementioned consist of millions of photos. Annotating photos or movies is therefore challenging work. Automated production of the image explanation may analyze every frame of the clip and then automatically produce appropriate textual explanation in order to efficiently and effectively finish the operation of video annotation. This considerably lessens the video worker's burden. Furthermore, annotation for photos and videos can help people with vision impairments comprehend a wide range of web videos and images.

Automated visual explanations are generated within the realms of clever surveillance, picture and video annotation, and human-computer interaction. This is only a small sample of the software used for captioning images. In conclusion, picture captioning has a wide range of applications that may improve labor efficiency and make life easier for individuals at work, home, and in the classroom.

8.2 Major challenges

Photo captioning research has evolved through multiple stages over an extended period, incorporating diverse technologies along the way. The use of NN engineering has created fresh opportunities in research generating captions for images, particularly in recent years. Even while neural networks' strong data processing abilities have produced some incredibly impressive results in the research of image caption generation, there are still certain issues that need to be resolved.

8.2.1 Image semantic richness

Although number of items in an image has no bearing on the current study's ability to describe image content, it can do so to some extent. The model often faces challenges in accurately describing objects, particularly when using concepts such as "two" or "group." Moreover, it tends to select multiple focal points in complex scenarios. Conversely, individuals can swiftly grasp the critical details of an image and absorb the relevant information. It won't be simple for the machine, though. Contemporary image captioning technology demonstrates greater proficiency in processing images with simple scenes. However, when confronted with complex landscapes and a multitude of object relationships, machines often encounter difficulties in accurately discerning the key contents of the image.

8.2.2 Inconsistent testing and training items

The present study states that the input for the network during training is combination of realistic terms and pictures, predicted word is the system's output. However, output of the network for each time step in the test method is the training dataset's word vector. The present method of training is largely influenced by choice of dataset. If a picture incorporates new things, method tries choosing nearest item using data set rather than original item. As a result, when the new objects are formed, a number of distinctions regarding training and testing procedure.

8.2.3 Multilingual textual description of pictures

A significant number of labelled training samples are required by the existing deep learning or machine learning approach for captioning photos. In practical applications, a written description of the image must be supplied in many languages to meet the demands of users with different local languages. Currently, writings produced in Chinese and English describe a large number of training instances however material presented in different languages contain few markups. If the textual descriptions of every language in the picture are completed, manual tagging would need great deal of effort and duration. Therefore, the integration of cross-language text descriptions for images in photo captioning represents a substantial issue and an area of exploration.

IX. Conclusion

Significant progress has been made in picture captioning in recent times. Latest developments in deep learning have increased the accuracy of picture captioning. Visual comprehension has a large variety of possible uses in fields like medical, security, and the military. By improving the efficacy of content-based picture retrieval, the image's written description can help. Moreover, the theoretical underpinnings and research methodologies of image scaptioning can propel advancements in various fields such as picture commenting, graphic quality assessment, cross-media search, clip subtitles, and dialogue in footage. These advancements hold significant applications in both academic and real-world contexts.

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)