Image Captioning Based on Deep Neural Networks

Asst. prof.- Prabhat Singh, Abhishek Pandey, Aditya Agarwal, Harshit Kumar Rai.
ABES Engineering College, *201009 Ghaziabad, Uttar Pradesh, India.*

Abstract. The fusion of computer vision and natural language processing has received a lot of interest recently thanks to the advent of deep learning. This field is represented by picture captioning, which trains a computer to comprehend an image's visual information using one or more phrases. The ability to analyse the state, the properties, and the relationship between these objects is also necessary for the meaningful description generating process of high level picture semantics. Despite the fact that image captioning is a challenging and intricate endeavour, numerous academics have made substantial advancements. In this research, we primarily discuss three deep neural network-based image captioning techniques: CNN-RNN based, CNN-CNN based, and reinforcement-based framework. Then, we briefly discuss the evaluation criteria, introduce the representative work for each of the three top approaches, and list the main advantages and difficulties.

1 Introduction

Computer vision in the field of image processing has significantly advanced in recent years, as evidenced by image classification [1] and object detection [2]. The issue of image captioning, which involves automatically generating one or more phrases to comprehend an image's visual information, has benefited from advancements in image categorization and object detection. Large-scale possible impacts of automatically creating complete and natural image descriptions include news image titles, medical image descriptions, text-based image retrieval, information accessed by blind users, and human-robot interaction. These captioning-related applications have significant theoretical and real-world research significance. Therefore, in the age of artificial intelligence, image captioning is a more challenging but important work.

An image captioning algorithm should produce a semantic description of a new image when given one. For instance, in Fig. 1, the waves, people, and boards make up the input image. An image's content is described in a sentence at the bottom. This sentence describes the scene, the action, and the items that are appearing in the image.

Humans can easily comprehend the content of an image and express it in the form of sentences using natural language, but for computers, the task of captioning images requires the integrated use of image processing, computer vision, natural language processing, and other significant areas of research findings. Designing a model that can fully utilise picture data to provide richer, more human-like image descriptions is the challenge of image captioning. In order to provide a meaningful description using high level picture semantics, it is necessary to be able to analyse the states of the objects or scenes in the image, comprehend their relationships, and produce a phrase that is both semantically and syntactically valid. Currently, it is unknown how the brain interprets.

Despite these obstacles, the issue has made substantial strides forward in recent years. There are commonly three types of image captioning algorithms. The first category, as depicted in Fig. 2. (a), approaches this issue utilising retrieval-based methods, which get the photos that are the closest matches first before transferring their descriptions to the query images' captions [3].
While these techniques can result in grammatically sound words, they are unable to change the captions to match the updated image. The second category in Fig. 2. (b),

A couple of surfers standing on their boards.
Figure 1 a sample caption for an image

which generates descriptions with predetermined syntactic rules and divides sentences into many pieces commonly employs template-based methods [4]. These techniques first use a number of classifiers to identify the items in an image, as well as their characteristics and relationships.

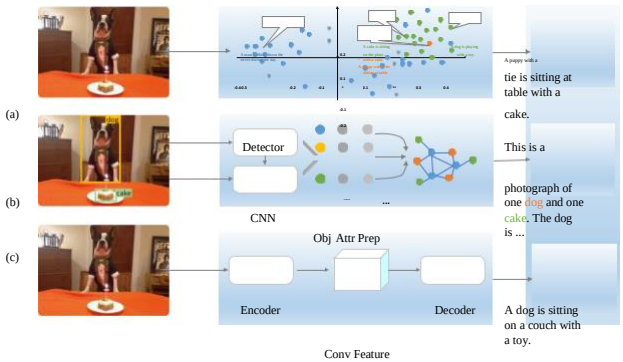

Figure 2 Three catogories for image captioning

Most recent works fall into the third category of neural network-based approaches in Fig2 due to the widespread use of deep learning (c). Most image captioning techniques today use a Convolutional Neural Network (CNN) as the encoder and a Recurrent Neural Network (RNN) as the decoder, particularly Long Short-Term Memory (LSTM) [6] to generate captions [7], with the goal of maximising the likelihood of a sentence given the visual features of an image. This approach is inspired by machine learning's encoder-decoder architecture [5]. CNN is sometimes used as the decoder while reinforcement learning is sometimes used as the decision-making network.

In this paper, we classify the neural network-based image captioning techniques into three groups based on the various encoding and decoding techniques: CNN-RNN based, CNN-CNN based, and reinforcement-based framework for image captioning. Their important points will be discussed in the section after this.

2 CNN-RNN based framework

An picture is composed of several colours in the human eye to represent various scenes. However, most images are created with pixels in three channels when seen through a computer. In contrast, various data modalities in the neural network are all heading toward creating a vector and performing the following operations on these features.

By embedding the input image into a fixed-length vector, it has been successfully demonstrated that CNNs can provide a rich representation of the image, which can then be used for a number of vision tasks such object recognition, detection, and segmentation [8]. As a result, CNN is frequently used as an image encoder in image captioning techniques based on encoder-decoder frameworks. The hidden layer, which has stronger training capabilities and can outperform mining deeper linguistic knowledge like semantic and grammatical information inherent in the word sequence, is continuously circulated by the RNN network to acquire historical data [9].

A recurrent neural network can be simply described in the hidden layer state for a dependence relationship between several location words in historical data. A CNN model for extracting image features serves as the encoder in an encoder-decoder process for captioning images. It can make use of models like ResNet [12], GoogleNet [11], VGG [10], and AlexNet [1]. The framework feeds the word vector expression into the RNN model in the decoder section. Each word is initially represented by a one-hot vector before becoming the same dimension as the image feature thanks to the word embedding model. The picture captioning issue can be expressed as a binary(I, S) problem, where I stands for a graph and S is a series of target words.

To address the issue of picture captioning, Mao et al. [13] created a multimodal Recurrent Neural Network(m-RNN) model that inventively blends the CNN and RNN models. The LSTM model is a unique sort of structure of the RNN model that can resolve the aforementioned issues because the gradient disappearance and limited memory problem of regular RNN. Three additional control units (cells), input, output, and forget gates, are added. The cells in the model will evaluate the information as it enters. Nonconforming material will be lost, while information that complies with the requirements will be preserved. The lengthy sequence dependency issue in the neural network can be resolved using this method. The NIC (Neural Image Caption) model was proposed by Vinyals et al. It uses an image as input in the encoder and uses LSTM networks in the decoder to produce the matching descriptions. The model effectively resolves the issue of vectorizing sentences in natural language. Using computers to process natural language is extremely important because it advances computer processing beyond the level of straightforward matching to that of semantic understanding.

The attention method in the field of computer vision is presented to encourage the alignment between words and image blocks. It is inspired by the neural network-based machine translation framework. So that the generated phrase is more in line with people's expressive habits, the "attention" transfer process of mimicking human eyesight can be mutually encouraged with the generation process of the word sequence. The attention mechanism adds the complete and spatial information pertaining to the image to the extraction of the image features, producing a richer statement description as opposed to recording the entire image as a static vector. Currently, the image characteristics are regarded as dynamic feature vectors paired with weights data.

The first attention mechanism was put forth in [15]. It suggested "soft attention," which involves choosing regions based on various weights, and "hard attention," which involves focusing attention on a specific visual concept. Deep neural networks with an attentional focus have produced impressive experimental outcomes. As seen in Fig.3, the model generates each word in accordance with the corresponding region of an image by means of an attention mechanism.

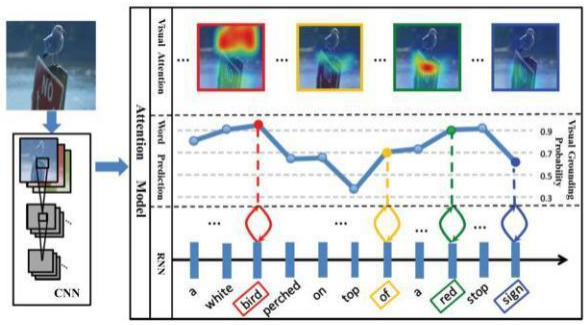

Figure 3 Illustration of the Attention Model

The CNN-based method makes use of convolutions. In contrast to the RNN, this component is feed-forward and lacks recurrent functionality. The CNN-CNN framework has a shorter training time per parameter, according to Aneja et al. [23], although the loss is larger for CNN than RNN. The CNN model's accuracy is due to the fact that Because of their less peaky word probability distributions, CNN is being penalised. Less peaky distributions, however, are not always a bad thing because they allow for various word predictions to be made for predicting a variety of captions, as seen in Fig. 4.

However, it also has two key flaws for the image captioning task, which encourage more in-depth study. The first is that there are differences between the measures utilised for training and testing. Metrics are non-differentiable and cannot be utilised as training loss directly, instead we use cross-entropy as the loss. Furthermore, while log probability may appear to give each word the same weight, in reality, people evaluate various words differently and give them varying weights. The "loss-evaluation mismatch" problem is the name given to this discrepancy [21]. The second is that each time step's input during training is taken from the actual caption, and each word that is formed is based on a prior word that was generated; When a word is not properly created.

3 CNN-CNN based framework

Despite the fact that models like LSTM networks feature memory cells that can retain the extensive history information of the sequence creation process better than RNN, it is still refreshed every time, making long-term memory somewhat challenging. Recent studies have demonstrated the advantages of CNN for picture captioning work, which were inspired by machine learning research. It has been shown to be extremely effective to use CNN in NLP for text production [22].
It has been demonstrated that the CNN convolution model can replace the RNN recurrent model in the field of neural machine translation because it not only outperforms the cycle model in terms of accuracy but also accelerates training time by a factor of nine. Since the translation work is done in a sequence-to-sequence architecture and in the task of creating image captions, an image is considered as a sentence in a source language, the majority of image captioning efforts are inspired by machine translation. To the best of our knowledge, Aneja et al[23] .'s work represents the first convolutional network for the text creation process in picture captioning, and we refer to this framework as CNN-CNN based.

Three key elements of this system resemble the RNN methodology. In both instances, the first and last components are word embeddings. However, masked LSTM or GRU (Gated Reccurent Unit) units are present in the centre component in the RNN situation.

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 triple gate of recurrence and the layered abstraction of convolution share a similar function. The goal is to ignore unimportant stuff and emphasise the relevant content, regardless of the methods used. As a result, there is little distinction between the convolutional model and the recurrent model in terms of accuracy. However, it is clear and undisputed that CNN training is faster than RNN training. There are two things that influence the unavoidable outcome.

Recurrent can only be handled sequentially, but convolutions can be processed in parallel. It is undoubtedly quicker to train parallel convolutional models on numerous machines than a serial recurrent model.

The convolution model's training can be accelerated using the GPU chip, although there is currently no

hardware to make RNN training faster.

In the 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 captioning. Since the CNN-CNN framework for image captioning was first put forth in 2017, machine translation has seen several advancements that can be applied to image captioning. More studies need to be done in the future to thoroughly examine CNN-CNN based attention mechanisms and the combination of CNN and RNN in the decoding stage.

4 Reinforcement based framework

Gaming, control theory, and other fields have all heavily utilised reinforcement learning. Determining an acceptable optimization goal for picture captioning is not as simple as it is for control or gaming problems, which by their very nature have real targets to improve.

When using reinforcement learning to caption images, the generative model (RNN) can be thought of as an agent that interacts with the outside world by taking input from the words and context vector at each time step. This agent's parameters specify a policy, whose execution causes the agent to choose a course of action. An action in the context of sequence generation is the prediction of the subsequent word at each time step. The agent (the hidden units of RNN) modifies its internal state after performing an action. The agent notices a reward once it has completed a sequence. The RNN decoder behaves like a stochastic strategy in such a setting, where selecting an action results in the generation of the subsequent word. By evaluating the series of actions from the current policy against the optimal action sequence, the PG method selects actions during training that are consistent with the current policy and only observes rewards at the conclusion of the sequence (or after the maximum sequence length). Finding the agent's parameters that maximise the anticipated reward is the aim of training.

The MIXER study [21], which compared the score of a candidate sentence to a reward signal in a reinforcement learning context, was the first to suggest the use of PG (policy gradient) to maximise non-differentiable objectives for image captioning. The MIXER method uses actions to train the RNN with the cross-entropy loss for several epochs using the ground truth sequences, which allows the model to focus on a good portion of the search space because the problem setting of text generation has a very large action space and makes the problem difficult to learn with an initial random policy. This innovative training method combines the reinforcement objective with the MLE (maximum likelihood estimation).

The visual semantic embedding that powers this reinforcement learning model outperforms other assessment measures without the need for retraining. Visual-semantic embedding, which measures the similarity between phrases and images, can also assess the accuracy of generated captions and serve as a good overall aim for image captioning optimization in reinforcement learning.

The decision-making network employs the "policy network" and the "value network" to jointly decide the next ideal word for each time step instead of learning the sequential loop model to identify the next correct phrase in a greedy manner. According to the present state, the policy network offers the confidence to forecast the following word. The reward value of each potential extension of the existing state is assessed by the value network.

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

Table 1 compares RNN, CNN, and Reinforcement Framework training parameters and training times (measured in seconds). The timings were acquired using a GPU from Nvidia Titan X. In comparison to the RNN and Reinforcement framework, we can train a CNN faster per parameter. However, CNN performs worse than the other models in terms of accuracy and diversity, as shown in the next section.

5 Evaluation metrics

The degree of similarity between the reference sentence and the caption sentence is primarily used in the current study to assess the advantages and disadvantages of the generating results. The five measuring indicators used most frequently are BLEU [16], METEOR [17], ROUGE [18], CIDEr [19], and SPICE [20]. 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.

The evaluation of image annotation results, which is based on n-gram precision, frequently uses BLEU. Calculating the separation between the evaluated and reference sentences is the basis of the BLEU measure. When the caption is closest to the reference statement's length, the BLEU approach tends to yield a better score.

An automated evaluation standard called ROUGE was created to assess text summarising algorithms. The ROUGE-N, ROUGE-L, and ROUGE-S evaluation criteria are three. ROUGE-N bases its evaluation on the given sentence, which computes a straightforward n-tuple recall for all reference statements: The largest common sequence (LCS), which calculates the recall, is the foundation of ROUGE-L. Based on the co-occurrence statistics of the skip-bigram between the reference text description and the prediction text description, ROUGE-S determines recall.

The harmonic mean of unigram precision and recall serves as the foundation for METEOR, however recall weights more heavily than accuracy. It differs from the BLEU in that it is present not only in the complete set but also at the sentence and segmentation levels and has a strong correlation with human judgement. It is also extremely relevant to human judgement.

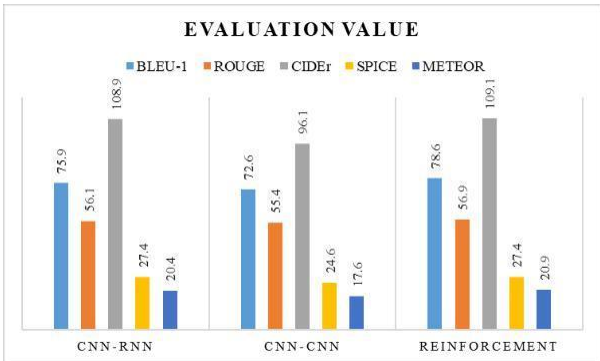

Figure 5 Evaluation Index of three methods

For five evaluation measures, we have displayed in Fig. 5 the best outcomes from the aforementioned three methodologies. We can observe that the CNN-RNN based and Reinforcement based approaches both outperform the CNN-CNN based framework in terms of performance, while maintaining a high level of accuracy. Additionally, the reinforcement structure performs the best because, as we discussed in Section 4, the objective function is more reasonable.

6 Discussions

6.1 Benefits

The ability to automatically annotate images could be useful from a theoretical and practical standpoint. The Internet's vast amount of data is the most significant factor in the current societal development process. The majority of these data are not standard data, and media data makes up a sizable component of them. They are frequently produced by online services like social networks or news outlets. Aside from the fact that humans can interpret these media images directly, there are limitations to the valuable data that machines can currently extract from them, and it is challenging to support humans in future work.

6.1.1 Intelligent monitoring

Intelligent monitoring enables the device to recognise and analyse the actions of individuals or cars in the captured scene, and under the right circumstances, it can produce alarms that will prompt the user to respond to emergencies and avert avoidable accidents. For instance, in channel monitoring, it gathers data on fairway operations and unlawful activities, keeps track of the fairway's state, and quickly learns about the usage of navigation channels, traffic patterns, and illicit sand mining. After then, report the issue to the command centre so that it can be scheduled, and immediately stop any illegal activity. This aspect can be addressed via image captioning.

6.1.2 Human-computer interaction

Robots are now used in an increasing number of industries due to the breakthroughs in science and technology as well as the necessity to develop human life. Robots with autopilot capabilities may intelligently avoid obstacles, change lanes, and pedestrians according on the road conditions they perceive. Driving can be done in a safe and effective manner, as well as automating tasks like parking. It can substantially improve people's lives and decrease safety accidents if the driver's hands and eyes are free. Better human interaction is required if the computer is to perform the task more effectively. Humans can use the machine's input to guide their processing after it informs them of what it observes.

6.1.3 Image and Video annotation

When a user submits an image, it must be labelled and annotated so that other users may quickly find it. The conventional approach is to find the closest comparable image in the database and get it for annotation, although this approach frequently yields wrongly tagged images. Additionally, video has developed into a need in people's life. Many movies now require subtitles in order to enjoy them more. Worldwide, a sizable amount of videos are made each year. Tens of thousands of images make up these videos. As a result, annotation of images and videos is a difficult undertaking. In order to efficiently and effectively complete the work of video annotation, the automatic generation of the picture description can process all the video frames and then automatically generate the corresponding text description in accordance with the content of each frame. This significantly reduces the workload of the video worker. Additionally, image and video annotation can assist those with visual impairments in understanding a huge variety of online movies and images.

In the areas of intelligent monitoring, human-computer interaction, and image and video annotation, the visual description is generated automatically. This is merely a portion of the image captioning software. In conclusion, image captioning can be used in many areas of people's lives, which can significantly increase labour productivity and ease people's daily lives, work, and learning.

6.2 Major challenges

The study of picture captioning has gone through a number of stages based on various technologies over a lengthy period of time. The use of neural network technology has created new opportunities for picture captioning research, 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.

6.2.1 Richness of image semantics

Although the 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. For instance, the model frequently struggles to adequately characterise the objects using concepts like "two" or "group." In addition, several focal locations are chosen in complex scenarios. People can quickly understand the image's crucial details and take in the pertinent information. It won't be simple for the machine, though. The current image description automatic generation technology can more fully describe images with basic scenes, but if the image contains complex scenes and a large number of item and object relationships, the machine frequently struggles to understand the image's key contents.

6.2.2 Inconsistent objects during training and testing

According to the current study, the network's output is the predicted word and its input during training is a real word vector or a combination of real words and images. The word vector in the training dataset's vocabulary is the output of the network at each time step in the test procedure, though. The choice of data sets has a significant role in the current training process. The method used is to choose the nearest object from the data set rather than the actual object when a given image contains novel things. As a result, when the new objects are formed, there are discrepancies in the training and testing process.

6.2.3 Cross-language text description of images

The current deep learning or machine learning method for captioning images needs a large number of labelled training samples. In order to satisfy the needs of users of various native languages, it is necessary in practical applications that a text description of the image be provided in a variety of languages. There are currently a lot of training samples described in texts written in English and Chinese, but there aren't many markups in texts written in other languages. Manual marking will take a lot of labour and time if the textual descriptions of each language in the image are done. Therefore, a major issue and a research challenge in picture captioning is how to incorporate cross-language text descriptions of images.

7 Conclusion

Recent years have seen considerable advancements in image captioning. Deep learning-based research from recent years has improved the accuracy of image captioning. The image's text description can increase the effectiveness of content-based image retrieval, broadening the spectrum of applications for visual comprehension in industries like medical, security, and the military, among others, with a wide range of potential uses. In addition, the theoretical underpinnings and research techniques of image captioning can support the theory and practise of image annotation, visual question answering (VQA), cross-media retrieval, video captioning, and video dialogue, all of which have significant academic and real-world 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. 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. 15. Xu, Kelvin, et al. "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention." Computer Science ,2048-2057. (2015)

16. 16. Papineni, K. "BLEU: a method for automatic evaluation of MT." (2001)

17. 17. Satanjeev, Banerjee. "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments." ACL-2005.228-231. (2005)