Image Captioning using Deep Learning

Rishi Kiran Reddy Bobbili [1], Anishkumar Rasik Dhamelia [2], Hitesh Kardam [3], Charles Quinn [4]

Stevens Institute of Technology

rbobbili@stevens.edu [1], adhameli@stevens.edu [2], hkardam@stevens.edu [3], cquinn3@stevens.edu [4]

Abstract

In recent years, deep learning techniques are being used in image recognition, due to their ability to identify and process images in detail. The objective of this study is to offer a thorough examination of various techniques used in image captioning. This includes the processes of visual encoding, text generation, training strategies, datasets, and evaluation metrics. To achieve this, we conduct a quantitative analysis of numerous state-of-the-art approaches, aiming to pinpoint the most influential advancements in architecture design and training strategies. Additionally, we delve into multiple variations of the problem and address the ongoing challenges it poses.

Introduction

Image captioning involves describing the visual content of an image using natural language. This requires a system that combines visual understanding with a language model capable of generating coherent and meaningful sentences. Recent advances in neuroscience have shed light on the connection between human vision and language generation [1]. Similarly, in the field of Artificial Intelligence, the development of architectures capable of processing images and producing language is a relatively recent focus of research. The aim of these research endeavors is to identify the most effective approach for processing an input image, representing its content, and converting it into a sequence of words. This involves establishing meaningful connections between visual and textual elements while ensuring the fluency and coherence of the generated language.

Over the past few years, the research community has made significant advancements in the design of image captioning models. Initially, deep learning-based approaches emerged, employing Recurrent Neural Networks (RNNs) that were fed with global image descriptors. Since then, the field has witnessed enrichments in methodologies, including the incorporation of attentive mechanisms, reinforcement learning techniques, as well as groundbreaking developments such as Transformers, self-attention, and single-stream BERT-like approaches.

Simultaneously, the Computer Vision and Natural Language Processing (NLP) communities have also made progress in addressing the challenge of establishing appropriate evaluation protocols and metrics for comparing image captioning results with human-generated ground truths. Despite the considerable investigation and improvements made in recent years, it is important to note that image captioning still remains an unsolved task [1].

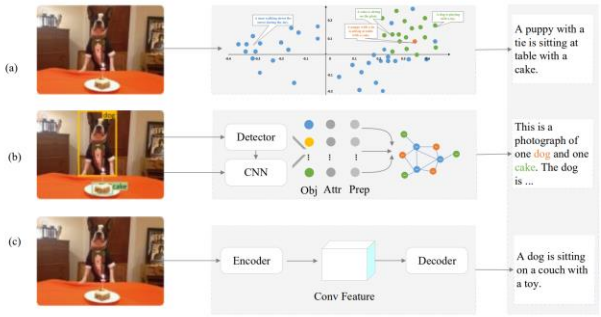

Figure 1: Three categories of image captioning

Despite the challenges faced, significant progress has been made in the field of image captioning over the past few years. Typically, image captioning algorithms can be classified into three categories. The first category, as depicted in Figure 1(a), employs retrieval-based methods. These methods retrieve the most similar images and transfer their de-

scriptions as captions for the query images. While these approaches can generate grammatically correct sentences, they lack the ability to adapt the captions to the specific characteristics of the new image.

The second category, shown in Figure 1(b), utilizes template-based methods. These approaches generate descriptions using predefined syntactic rules and often break sentences into multiple parts. With the widespread adoption of deep learning techniques, the majority of recent works fall into the third category, illustrated in Figure 1(c), known as neural network-based methods. Drawing inspiration from the encoder-decoder architecture in machine learning, these methods employ a Convolutional Neural Network (CNN) as the encoder and a Recurrent Neural Network (RNN), particularly Long Short-Term Memory (LSTM), as the decoder. The objective is to maximize the likelihood of a sentence given the visual features of an image.

In this paper, we categorize the neural network-based image captioning methods into three groups based on their encoding and decoding techniques: CNN-RNN based, CNN-CNN based, and reinforcement-based frameworks. In the following section, we will discuss their main concepts and approaches in more detail [10].

To sum up, this paper will focus on:

- Effectiveness of CNN-RNN over other methods for image captioning.
- Explore variants of RNN like long short-term memory networks (LSTM).
- Challenges during evaluating image-captioning.
- Comparing our models with previous works.
- Future expansion into other fields (Videos, Augmented reality etc.).

CNN-CNN Framework

Despite the advantages of LSTM networks, which can effectively retain long-term information during sequence generation, they still face challenges in updating their memory cells at each time step. This limitation makes it difficult to maintain long-term memory. However, recent studies have drawn inspiration from machine learning and demonstrated the advantages of using Convolutional Neural Networks (CNN) in image captioning. The application of CNN in Natural Language Processing (NLP) for text generation has been shown to be highly effective. In the field of neural machine translation, it has been proven that replacing the recurrent model (RNN) with a convolutional model (CNN) not only improves accuracy compared to the cyclic model but also significantly speeds up training by a factor of nine.

Many image captioning approaches draw inspiration from machine translation, as both tasks involve sequence-to-sequence architectures. In image captioning, an image is treated as a sentence in a source language. To the best of our knowledge, the first utilization of a convolutional network for the text generation process in image captioning was introduced by Aneja et al. [11], which we refer to as the CNN-CNN based framework.

Reinforcement based framework

Reinforcement learning has found widespread applications in various fields such as gaming and control theory. In control or gaming problems, there are specific objectives that can be optimized naturally. However, defining an appropriate optimization goal for image captioning is a complex task.

When integrating reinforcement learning into image captioning, the generative model (RNN) can be considered as an agent that interacts with the external environment, which consists of the words and the context vector serving as inputs at each time step. The agent's parameters define a policy, which determines the action taken by the agent. In the context of sequence generation, an action refers to predicting the next word in the sequence at each time step. After taking an action, the agent updates its internal state, represented by the hidden units of the RNN. Upon reaching the end of a sequence, the agent receives a reward.

In this framework, the RNN decoder functions as a stochastic policy, where selecting an action corresponds to generating the next word. During training, the policy gradient method selects actions based on the current policy and only receives a reward at the end of the sequence (or after reaching the maximum sequence length). The training process involves comparing the sequence of actions generated by the current policy against the optimal action sequence. The objective of training is to find the agent's parameters that maximize the expected reward.

CNN-RNN Framework

From a human perspective, an image is perceived as a composition of different colors, forming various scenes. However, computers typically represent images using pixels in three channels. Nevertheless, in the realm of neural networks, different types of data are converging towards the creation of vectors, allowing subsequent operations on these features. Convincing evidence has demonstrated that Convolutional Neural Networks (CNNs) can generate a comprehensive representation of an input image by embedding it into a fixed-length vector. This representation can then be utilized for diverse vision tasks, including object recognition, detection, and segmentation.

Therefore, in image captioning methods that employ encoder-decoder frameworks, it is common to employ a CNN as the image encoder. On the other hand, Recurrent Neural

Networks (RNNs) capture historical information by propagating hidden layer outputs in a cyclic manner. RNNs possess robust training capabilities and can outperform deeper linguistic knowledge extraction, such as mining the semantic and syntactic information implicitly present in a sequence of words.

Mao et al. [12] proposed a multimodal Recurrent Neural Network(m-RNN) model that creatively combines the CNN and RNN model to solve the image captioning problem. Because of the gradient disappearance and the limited memory problem of ordinary RNN, the LSTM model is a special type of structure of the RNN model that can solve the above problems. It adds three control units (cell), which are the input, output and forgot gates. As the information enters the model, the information will be judged by the cells. Information that meets the rules will be left, and nonconforming information will be forgotten. In this principle, the long sequence dependency problem in the neural network can be solved [10].

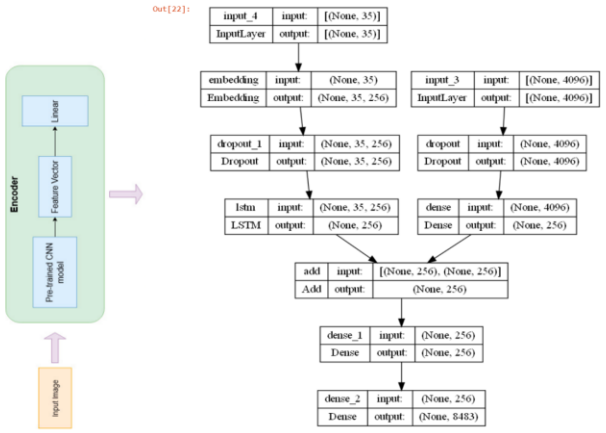

Figure 2: CNN-RNN

Reasons to choose CNN-RNN

In this project, we have chosen to utilize the CNN-RNN framework due to its overall efficiency compared to other frameworks such as CNN-CNN and reinforcement learning-based approaches. Our decision is supported by previous research that has conducted a comparative analysis of these three techniques.

The paper titled "Image Captioning Based on Deep Neural Networks" authored by Shuang Liu, Liang Bai, Yanli Hu, and Haoran Wang [10] provides valuable insights into the performance of these techniques on the COCO dataset. By examining their findings, we can draw meaningful conclusions and make an informed selection of the CNN-RNN framework for our project.

Method	Parameters	Time/Epoch
CNN-RNN	13M	1529s
CNN-CNN	19M	1585s
Reinforcement	14M	3930s

Figure 3: Training time comparison

The training time comparison of all three frameworks on a single dataset is illustrated in Figure 3. The results demonstrate that both the CNN-CNN and CNN-RNN frameworks exhibit faster training times compared to the Reinforcement framework.

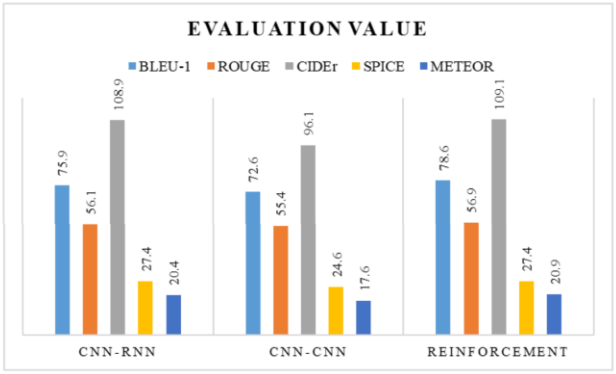

Figure 4: Evaluation of the three methods

Figure 4 presents the top-performing results of the three methods mentioned earlier for five evaluation metrics. From the figure, it is evident that both the CNN-RNN-based and Reinforcement-based methods outperform the CNN-CNN-based framework in terms of performance across all five metrics.

By examining the overlapping results of the above comparisons, we can draw a conclusive inference that the CNN-RNN framework not only demonstrates faster training times but also delivers superior performance in terms of results making it the ideal choice for this project.

Dataset

This project used Flicker8k and Flicker30k. The Flickr8k dataset has played a significant role in advancing image captioning research. With its diverse collection of 8,000 images and their corresponding five captions, the dataset provides ample training and evaluation data for developing and benchmarking image captioning models. These captions are written in natural language, making the dataset an ideal resource for training models to generate human-like descriptions for visual content.

While newer datasets such as COCO (Common Objects in Context) and Conceptual Captions have emerged as popular benchmarks due to their larger size and broader coverage,

the Flickr8k dataset continues to be widely used and referenced in the research community. Its established reputation, accessibility, and well-defined annotations make it a reliable resource for evaluating the performance of image captioning algorithms and comparing them against previous work.
The availability of the Flickr8k dataset for research purposes, along with its downloadability from the official website, facilitates reproducibility and fosters collaboration among researchers in the field of image captioning. By leveraging this valuable dataset, researchers can continue to push the boundaries of image understanding and language generation, ultimately contributing to the advancement of image captioning technologies.

Evaluation Metric

Assessing the quality of generated captions is a challenging and subjective task, particularly because captions must not only be grammatically correct and fluent but also accurately describe the input image. While human evaluation is considered the most reliable method, it is costly and lacks reproducibility, making it difficult to compare different approaches on a fair basis. To address this, automatic scoring methods have been developed, which typically involve comparing system-generated captions with human-produced reference sentences. However, some metrics do not rely on reference captions and offer alternative approaches for evaluating caption quality.

BLEU Score

To evaluate the model's captions, we calculated their BLEU (Bilingual Evaluation Understudy) scores. BLEU scores are a popular metric used to evaluate the quality of NLP models, particularly in tasks such as machine translation, image captioning, text summarization, and language generation. When it comes to evaluating the output of NLP applications where the output is a sentence, determining how "good" the output becomes challenging due to the lack of a single correct answer. BLEU scores provide a quantitative measure to assess the similarity between the predicted output and human-generated target sentences.
The underlying idea behind BLEU scores is to compare the predicted sentence with one or more reference sentences and measure their overlap in terms of n-grams. An n-gram refers to a sequence of 'n' consecutive words in a sentence. BLEU scores consider different n-gram precisions (1-gram, 2-gram, 3-gram, etc.) to capture the accuracy of predictions at different levels of word sequences.
To calculate BLEU scores, precision is computed for each n-gram, which represents the number of correctly predicted n-grams divided by the total number of predicted n-grams. However, to address the issues of repetition and multiple target sentences, a modified precision called "clipped precision" is used. Clipped precision compares each predicted word with the target sentences and limits the count for each correct word to the maximum count in any target sentence. This approach avoids overestimating precision due to repetition or considering only a single correct target sentence. Additionally, a "brevity penalty" is applied to avoid favoring short predicted sentences.
BLEU scores range between 0 and 1, where 1 represents a perfect match between the predicted and target sentences. However, achieving a score close to 1 is unrealistic in practice because different variations of correct sentences exist. Scores around 0.6 or 0.7 are considered good and indicate reasonably accurate predictions [7].

Our BLEU Scores

Our model's scores fell near this range for unigrams and bigrams, but performed worse for 3-grams and 4-grams.
While BLEU scores have their strengths, such as being quick to compute, language-independent, and widely used for comparison, they have some shortcomings. BLEU scores solely focus on exact word matches and do not consider word meaning, variations, or word order. This is evident in some of our results shown in figure 5.

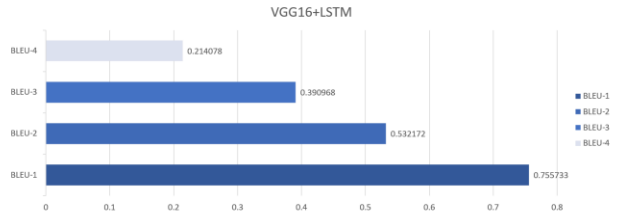

Figure 5: Our results

Sample Outputs

Figure 6

For example, the generated caption for the image in figure 6 fails to preserve the exact relationship between the elements in the scene. The generated caption describes the girl as "sitting in front of a rainbow-colored bowl," which is not the true context. The girl is sitting in front of a painted rainbow, not a bowl. And the bowl itself is not rainbow-colored. Additionally, the generated caption fails to mention "paint," which is a prominent element of the image and is mentioned in all of the sample captions.

Even though the generated caption is not entirely accurate, it receives a relatively high BLEU score since it includes 5-gram matches with the sample captions like "sitting in front of rainbow" and "is sitting in front of."

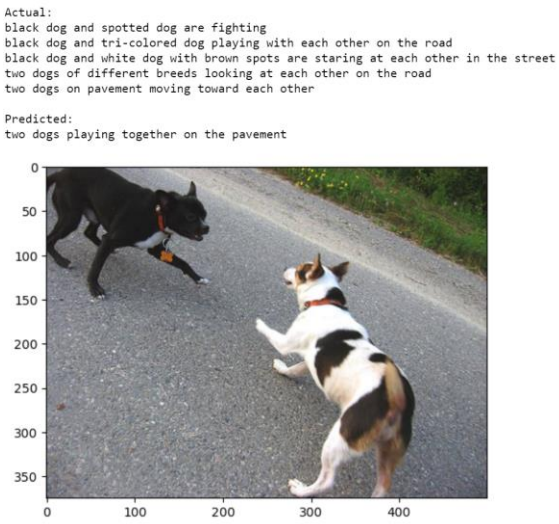

Figure 7

The next example, shown in figure 7, is of one of the more successful captions. The generated caption appears to naturally and accurately describe the image. However, it does include slightly less detail than the sample captions. Four of the five example captions describe that the two dogs are of different breeds or have different color fur. The generated caption omits these details about the dogs.

Despite these shortcomings, the generated captions still exhibit a level of appropriateness and human-like expression. They generally capture the main elements of the images and provide a coherent description. The model has successfully learned to generate captions that align with the overall structure and style of the training data. However, there is room for improvement in terms of contextual relationships and incorporating finer detail.

Conclusion

In summary, the objective of this image captioning project was to present a comprehensive overview of image captioning techniques, covering visual encoding, text generation, training strategies, datasets, and evaluation metrics. Through a comparative analysis of various state-of-the-art approaches, we identified significant advancements in architecture and training strategies that have made an impact in the field.

After careful consideration, we selected the CNN-RNN framework for our project due to its efficiency in terms of training time and its ability to deliver superior performance when compared to the CNN-CNN and reinforcement-based frameworks.

To assess the quality of the generated captions, we employed BLEU scores, a widely used metric in NLP, which quantifies the similarity between predicted captions and reference sentences.

The implementation of the CNN-RNN framework on the Flicker8k database was successfully accomplished, yielding satisfactory results. The Python programming language, along with the TensorFlow library, was utilized for testing and validating our framework.

Overall, this project makes a valuable contribution to the advancement of image captioning research by offering insights into different approaches, conducting a comparative analysis, and employing appropriate evaluation metrics. The findings and methodologies presented here can serve as a guide for future research and development, leading to enhanced image captioning systems with practical applications.

Future Research

The field of image captioning is continually evolving, and there are several potential directions for future expansions and advancements. Here are some key areas that researchers and developers are exploring:

1) Video Captioning: is the task of automatically generating textual descriptions or captions for videos. It involves analyzing the visual content of a video and generating a coherent and semantically meaningful description that accurately represents the events, actions, and objects depicted in the video.

2) Augmented Reality (AR) technology, combined with image captioning, offers exciting possibilities for enhancing user experiences and interactions with the physical world.

3) Image captioning with voice recognition combines the capabilities of image understanding and speech recognition to generate textual descriptions of images based on spoken commands or queries.

References

[1] From Show to Tell: A Survey on Deep Learning-based Image Captioning. Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara.

[2] J.-Y. Pan, H.-J. Yang, P. Duygulu, and C. Faloutsos, "Automatic image captioning," in ICME, 2004.

[3] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, "Every picture tells a story: Generating sentences from images," in ECCV, 2010.

[4] V. Ordonez, G. Kulkarni, and T. Berg, "Im2text: Describing images using 1 million captioned photographs," in NeurIPS, 2011.

[5] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "DeViSE: a deep visual-semantic embedding model," in NeurIPS, 2013.

[6] Hu, Z.; Zhao, Y.; Khushi, M. A Survey of Forex and Stock Price Prediction Using Deep Learning. Appl. Syst. Innov. 2021, 4, 9.

[7] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in ACL, 2002.

[8] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in ACL Workshops, 2005.

[9] Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. Laith Alzubaidi1, Jinglan Zhang, Amjad J. Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, J. Santamaría, Mohammed A. Fadhel, Muthana Al-Amidie and Laith Farhan.

[10] Image Captioning Based on Deep Neural Networks: Shuang Liu, Liang Bai, Yanli Hu and Haoran Wang.

[11] Aneja, Jyoti, A. Deshpande, and A. Schwing. "Convolutional Image Captioning." (2017).

[12] Mao, Junhua, et al. "Explain Images with Multimodal Recurrent Neural Networks." Computer Science (2014)