Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

Advancements in Deep Learning-Based Image Captioning

Siqi Wang [1], Jihong Zhuang [2, *]

[1] School of Mechanical, Electrical and Information Engineering, Xiamen Institute of Technology, Xiamen, China

[2] School of Automation Science and Engineering, South China University of Technology, Guangzhou, China

* Corresponding Author Email: 202130462363@mail.scut.edu.cn

Abstract. In the confluence of natural language processing and machine vision, the field of image captioning has experienced exponential growth since the introduction of the BLEU evaluation algorithm by IBM in 2002. This discipline serves to bridge the "semantic gap" between human and machine perception, translating visual information into semantic narratives. Such technology is extensively applied in areas like human-computer interaction, video subtitling, quiz generation, and image-based search functionalities. The paper presents an analysis of two primary methodologies in image captioning: template-based and encoder-decoder-based structures. Template-based approaches, defined by the use of pre-set templates, ensure syntactic accuracy yet offer limited flexibility in caption generation. Innovations within this methodology, including paraphrase back-translation and the integration of psycholinguistics, have enhanced caption diversity and descriptiveness. On the other hand, the encoder-decoder framework, particularly the CNN-RNN model, utilizes deep neural networks to learn directly from image-caption pairs. This method represents a more dynamic and adaptable approach to caption generation. The amalgamation of Convolutional Neural Networks (CNN) with Long Short-Term Memory (LSTM) networks within this framework has notably advanced the descriptive quality of captions, effectively addressing complex image contexts.

Keywords: Image captioning; deep learning; template-based structure; encoder decoder-based structure.

1. Introduction

Since its proposal by IBM in 2002, the BLEU algorithm has been pivotal in evaluating machine translation tasks, marking a significant convergence of natural language processing and machine vision [1]. This blend, driven by deep learning, aims to develop algorithms enabling computers to extract information about objects in images, including their locations, and to generate captions by integrating semantic and visual features. This technology finds widespread application in areas such as human-computer interaction, video subtitling, video quizzes, and image-based search.

Research in image caption generation continues to evolve, leading to an array of new algorithms and models. These methods generally fall into two categories [2]: Template-based methods: This approach utilizes templates with placeholders for generating captions. It identifies various elements, attributes, and behaviors, which are then used to fill these placeholders. While template-based methods can produce captions free from syntactic errors, their reliance on preset templates limits their ability to create captions of variable lengths. Encoder-decoder structure-based approaches: Advancements in deep neural networks have led to the adoption of CNNs and RNNs for tasks like machine translation. These approaches involve training models using images and corresponding captions, aiming to maximize the likelihood function. This method allows for direct learning of all parameters during training [3].

This paper presents a detailed analysis and review of the research developments and current status of image caption generation algorithms based on deep learning. It includes an introduction to the

algorithms within these classifications, their experimental outcomes, and projections for future developments in this field.

2. Related to the topic

2.1. Definition of Image Captioning Generation

Humans can establish relationships based on the visual information in images, and then perceive the high-level semantic information of images; however, computers can only extract the feature information of images, and cannot generate high-level semantic information as human brains do, which is the problem of "semantic gap" between humans and computers [4]. Image captioning technology can transform visual information into semantic information, which makes up for this gap and helps to solve the "semantic gap".

Image captioning was born in 2002 after IBM proposed the use of BLEU as an evaluation algorithm for machine translation tasks, and after more than 20 years of development, the technology can maintain a certain accuracy rate for image description generation, and has been developing rapidly in the direction of meeting the needs of diversified markets while continuously improving the accuracy rate.

2.2. The role of deep learning in research

Deep learning has greatly impacted the field of image captioning generation. When performing image description tasks, deep learning learning models typically extract image features through Convolutional Neural Networks (CNN) and simultaneously utilize models such as Recurrent Neural Networks (RNN) to generate natural language descriptions [5]. Deep learning has had a profound impact on the field of image captioning.

Learning to generate descriptions directly from raw image data realizes end-to-end image annotation, and by training large-scale datasets, these models can generate rich, specific image descriptions covering object recognition, scene description, action inference, etc. Simplifying the process of image annotation reduces the difficulty of model training, and greatly improves the quality and diversity of image description generation [6]. This is of great significance for image understanding, intelligent search, and assisting the visually impaired.

3. Methods

3.1. Method based on template structure

The realization of this method is based on manually set various types of templates, so the process requires more manual participation, the general operation process: first of all, through the classifier to obtain the object feature information in the image, and then through the templates set up in advance will be the information corresponding to the generation of text. The template structure-based method is shown in Fig. 1 [7].

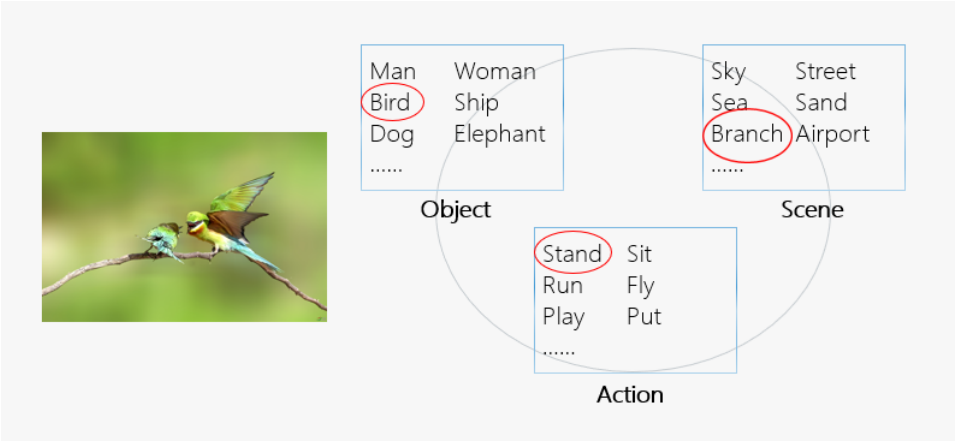

Fig. 1 Template-based method (Photo/Picture credit: Original).

In recent years, in the field of image captioning generation, there are numerous researchers using template class-based methods to make results toward diversified goals [8]. Ingrid Ravn Turkerud and others added captions to the datasets by back-translating the paraphrase to augment the original datasets and improve the performance of image captioning model; Kazuki Umemura and others use a method that calculates sentence imaginability scores based on word-level imagery scores (New captioning methodology incorporating the psycholinguistics concept of imaginability)By incorporating the imaginability of image content, the method aims to generate more vivid and descriptive captions that capture the visual content; Xianyu Chen and others proposed a self-distillation approach to solve the problem of image captioning generation in the case of fewer samples [9]. The technique of training another model by using the captions generated by the model to improve its performance and generalization abilit; Andrej Karpathy and others introduced a multimodal embedding space that maps the visual features and textual descriptions of an image into a common multimodal embedding space, which makes it possible to achieve semantic alignment between image and text, and to automatically generate natural language descriptions matching the content of the image.

In summary there is a wide variety of methods based on the template class, which can be used to design lexical model templates according to the different needs of the operator, as well as to change the way of extracting image data [10].

3.2. Method based on encoder decoder structure

The CNN-RNN model is one of the most classical encoder-decoder approaches. An encoder is usually a convolutional neural network (CNN), which usually consists of a series of convolutional and pooling layers. A pre-trained CNN is used to extract high-level features of the image. Through multiple convolution and pooling operations, the encoder gradually reduces the spatial dimensions of the image and extracts high-level abstract features. These feature vectors contain high-level semantic information about the image. These features are later used as the basis for the decoder to generate text descriptions; the decoder is usually a Recurrent Neural Network (RNN), which is capable of generating sequences of words step-by-step, taking the hidden state of the previous time step as input while generating each word, taking into account the previous text information. During training, the model is asked to generate the correct captioning for each image as accurately as possible. It is common to define a loss function (cross-entropy loss function) and optimize the model parameters by maximizing the likelihood probability to minimize the loss function so that image captioning generation are as close as possible to the true descriptions, thus improving the model's fit to the observed data.

CNN-LSTM model is similar to CNN-RNN, this approach also uses convolutional neural network to extract image features but the decoder part uses Long Short Term Memory Network (LSTM) to generate text description. Using Long Short-Term Memory Network (LSTM), the image feature

vectors generated from the encoder are received and words or phrases describing the content of the image are generated one by one. At each step of the generation process, the LSTM considers previously generated words and current image features as a way to predict the next most likely word, and eventually LSTM generates the corresponding textual descriptions by continuously combining the vectorized image feature information with the semantic information. However, most encoder-decoder models require a fixed-length output sequence during training, which may limit the model from generating flexible and variable-length descriptions. For image content of varying length and complexity, a fixed-length output sequence may not be sufficiently expressive.

The CNN-Transformer model combines a convolutional neural network and a Transformer model, where a CNN is used in the encoder part to extract image features and a Transformer is used in the decoder part to generate descriptive text. Typically, pre-trained convolutional neural networks (e.g., ResNet, VGG, etc.) can be used to extract image features that contain global and local information in the image. The extracted image features will be passed to the Transformer decoder. The Transformer model is able to effectively model long-range dependencies and also has the advantage of parallel computation, which makes it perform well in processing sequential data. In the image caption generation task, the Transformer decoder will utilize the image feature vectors from the CNN encoder and generate descriptive text step by step. Since the Transformer model's self-attention mechanism can better capture the dependencies between sequences and the global context, the CNN-Transformer model may show better performance than the traditional CNN-RNN model in the image subtitle generation task in some cases, especially when dealing with long text descriptions and complex contexts. That is, the Transformer model performs excellently when dealing with long distance dependencies. The encoder decoder structure-based method is shown in Fig. 2.

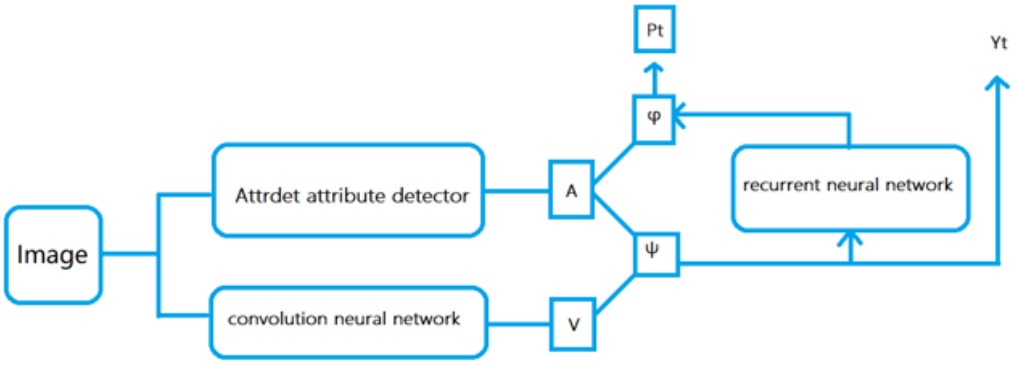

Fig. 2 Encoder decoder structure-based method (Photo/Picture credit: Original).

The Encoder-Decoder with Attention model is a commonly used approach that combines an encoder, decoder, and attention mechanism for more accurate and smooth image captioning generation. In this approach, the encoder typically uses a Convolutional Neural Network (CNN) to extract a feature representation of the input image. These feature vectors capture the important information in the image, which is then passed to the decoder for generating descriptive text. The decoder part usually uses variants of Recurrent Neural Networks (RNN) such as Long Short-Term Memory Networks (LSTMs) or Gated Recurrent Units (GRUs) to deal with tasks that do not involve long sequences of indeterminate data, such as generating image captioning. Unlike the traditional encoder-decoder structure, the Encoder-Decoder with Attention model introduces an attention mechanism. As the decoder generates each word or phrase, the attention mechanism allows the model to dynamically focus on different regions of the image encoder, thus enabling the model to better capture the correspondence between the image and the text, and improve the accuracy and fluency of the generated descriptions. This attention mechanism helps the model to pay more attention to the image regions that are relevant to the currently generated word, thus making the generated text more descriptive. The topic of Jingqiang Chen, Hai Zhuge's research is to add captions to news images, while News Image Captioning contains more detailed information than general image captioning, such as entity names and events, they generate captions for news images through a multimodal attention encoder-decoder model. Their research methodology consists of four main parts: the text

encoder encodes the text through an RNN model, and the image encoder completes the encoding of the image by obtaining the vector representation of the image through Oxford VGGNet. The decoder decodes the words through the RNI model. In order to balance the text encoding and image encoding, they developed a multimodal attention mechanism based on the traditional attention mechanism. In their experimental data analysis, it can be seen that NNattSim, which combines text and image encoding, outperforms NNattImg, which only pays attention to the image input, and NNattTxt, which focuses on the text summarization problem. Umemura, K et al. They also used the attention mechanism to connect the encoder to the decoder, where the encoder is used to convert the image into a vector and the decoder is used to generate the text, which can be used to generate an accurate description of the image based on the output of the encoder. The concept of imageability is introduced and incorporated into the encoder-decoder architecture. By considering image imaginability, this method can generate more vivid and specific image captions, through which the quality and accuracy of image captions can be improved. Overall, this model introduces an attention mechanism that allows the decoder to weight attention to different parts of the encoder in generating each word in order to improve the accuracy and coherence of the description.

Pre-trained Vision-and-Language Models for the task of image caption generation is a state-of-the-art approach. These models enable a better understanding of the relationship between images and semantics by jointly learning image and text representations, leading to more accurate and semantically rich image caption generation [10]. Models are typically based on the Transformer architecture and are pre-trained with large-scale datasets to learn joint representations of images and language. The pre-training phase of its methodology accomplishes the image-text pairing task by way of modeling of the model, such as image caption generation or image quizzing. In this way, the model learns the connection between visual features, semantic understanding and language generation. In the application phase, the pre-trained model can be used as a decoder for image subtitle generation. For an input image, the model is able to apply the attention mechanism to focus on various parts of the image. Combing image features with previously generated text content to generate the next word or phrase. It is possible to generate image descriptions that are relevant to the image content and are more accurate and semantically rich.

The encoder-decoder architecture can achieve end-to-end learning from input data to output data without having to manually design the feature extractor. It realizes effective conversion between image and language, improves the model's ability to understand and describe image content, and provides a powerful framework for multimodal information processing.

4. Challenges

Current encoder-decoder-based methods exhibit instability in handling interpretation due to fixed-length output and localized attention. Similarly, template-based approaches face constraints, notably in producing uniform, static content, coupled with a high degree of required human intervention. Given the escalating market demands, the field of image captioning generation continues to present extensive opportunities for advancement.

5. Conclusion

This manuscript presents a comprehensive survey of ongoing research in the field of image caption generation. It outlines the evolution and present state of this technology, detailing the nuances of two primary methodologies: template-based and encoder-translator-based approaches, through a methodical categorization. Looking ahead, it is acknowledged that market demands are varied, necessitating the enhancement of the model's adaptability to produce tailored descriptions pertinent to distinct application contexts. Additionally, there is a need for refinement of the evaluation algorithms to elevate the quality of assessments of the model's outputs and to facilitate continual refinement in the comparison of these outputs. Addressing the influence of external disturbances on the model's output is also crucial for augmenting its stability.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

Reference

[1] Mahalakshmi, P., & Fatima, N. S. (2022). Summarization of text and image captioning in information retrieval using deep learning techniques. IEEE Access, 10, 18289-18297. Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pp. 311-318.

[2] Ma, H., Zhu, J., Lyu, M. R. T., & King, I. (2010). Bridging the semantic gap between image contents and tags. IEEE Transactions on Multimedia, 12(5), 462-473.

[3] Turkerud, I. R., & Mengshoel, O. J. (2021, December). Image captioning using deep learning: text augmentation by paraphrasing via backtranslation. In 2021 IEEE Symposium Series on Computational Intelligence (SSCI) (pp. 01-10). IEEE.

[4] Umemura, K., Kastner, M. A., Ide, I., Kawanishi, Y., Hirayama, T., Doman, K., ... & Murase, H. (2021). Tell as you imagine: Sentence imageability-aware image captioning. In MultiMedia Modeling: 27th International Conference, MMM 2021, Prague, Czech Republic, June 22–24, 2021, Proceedings, Part II 27 (pp. 62-73). Springer International Publishing.

[5] Chen, X., Jiang, M., & Zhao, Q. (2021). Self-distillation for few-shot image captioning. In Proceedings of the IEEE/CVF winter conference on applications of computer vision (pp. 545-555).

[6] Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3128-3137).

[7] Chen, J., & Zhuge, H. (2019, September). News image captioning based on text summarization using image as query. In 2019 15th International Conference on Semantics, Knowledge and Grids (SKG) (pp. 123-126). IEEE.

[8] Umemura, K., Kastner, M. A., Ide, I., Kawanishi, Y., Hirayama, T., Doman, K., ... & Murase, H. (2021). Tell as you imagine: Sentence imageability-aware image captioning. In MultiMedia Modeling: 27th International Conference, MMM 2021, Prague, Czech Republic, June 22–24, 2021, Proceedings, Part II 27 (pp. 62-73). Springer International Publishing.

[9] Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J., & GAO, J. (2020, April). Unified vision-language pre-training for image captioning and vqa. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 07, pp. 13041-13049).

[10] Zeng, Y., Zhang, X., Li, H., Wang, J., Zhang, J., & Zhou, W. (2023). X 2-vlm: All-in-one pre-trained model for vision-language tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence.