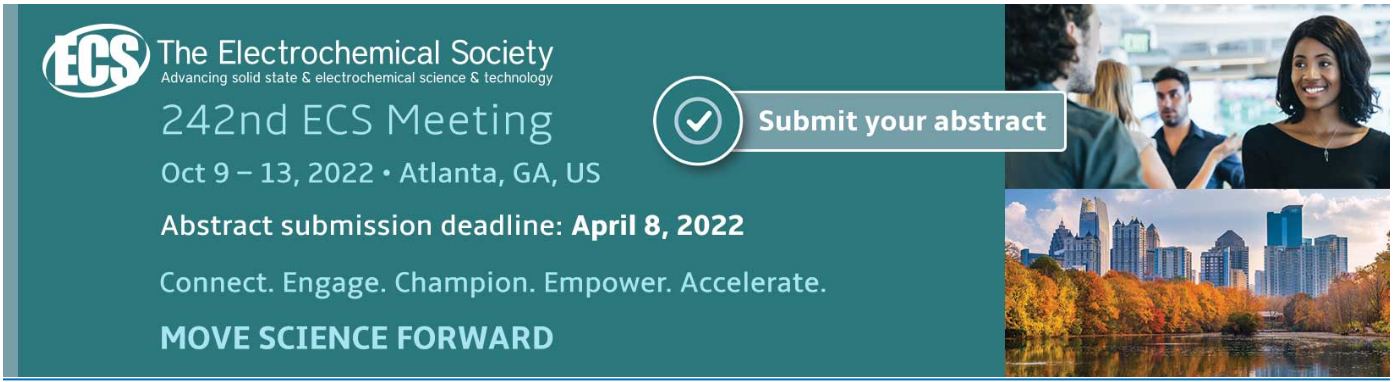
Journal of Physics: Conference Series

PAPER • **OPEN ACCESS**

An Integrative Review of Image Captioning Research

To cite this article: Chaoyang Wang *et al* 2021 *J. Phys.: Conf. Ser.* **1748** 042060

View the article online for updates and enhancements.



An Integrative Review of Image Captioning Research

Chaoyang Wang[1], Ziwei Zhou[1*] and Liang Xu[1]

[1] School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, Liaoning, 114051, China

[*]Corresponding author's e-mail: zzw@ustl.edu.cn

Abstract. In the field of computer vision, image captioning is a new frontier research. The basic task of image captioning is to generate a descriptive natural language for the input image. This paper investigates and analyzes the related research of image captioning. Firstly, the task and application scenarios of image captioning are introduced. Secondly, the image captioning algorithm based on template and the image captioning algorithm based on encoder-decoder structure are analyzed, and the advantages and limitations of each method are discussed. Then, the benchmark dataset and evaluation for image captioning are introduced. Finally, the future development of image captioning is prospected.

1. Introduction

The acquisition and analysis of visual information is one of the ways for people to understand the world. The scene pictures can be transformed into their own cognition. Through continuous accumulation, people can gradually understand the surrounding world. Image captioning is a research direction that transforms images into cognition, Its basic model needs to have two parts of functions. The first part is the extraction of image feature information, which mainly needs to extract the object information and object location information in the image; the second part is to analyze the semantic information of image description and combine it with image features to generate image description. The human brain has a complete cognitive system. As long as the image is received, the brain will process the image and analyze the image content. When the computer realizes image captioning, it is often necessary to instill the ability of cognitive image into the computer. The traditional program can not realize this function. On the one hand, the logic to be considered is too complex and the program is too large; on the other hand, the traditional program is too rigid to achieve the expected effect. Driven by the background of artificial intelligence, the algorithm of neural network is adopted to make the computer more close to people's cognitive ability and reach the language description level of children.

Image captioning is one of the hot issues in the field of artificial intelligence, It has a wide range of application scenarios, It can be used in human-computer interaction, adding subtitles to video [1], video question answering [2], search important information according to image content and image search by keywords, etc. As many as 17.31 million people are visually impaired in China. Their travel is very inconvenient and it is difficult to avoid road emergencies in time, the application of image captioning in road condition recognition enables the visually impaired to perceive the external environment in real time, which provides great convenience and safety for travel.

The organizational structure of this paper is as follows. The second part mainly introduces and analyzes the common network model of image captioning and its design ideas. The third part mainly introduces the commonly used data sets and performance evaluation indexes of image captioning training model. The fourth part summarizes the current status of image captioning, and puts forward the future development prospects of this direction.

2. Image Captioning Method

Image captioning algorithms are mainly divided into two types, one is based on the template method; the other is based on the encoder decoder structure. The following mainly analyzes the two methods. The development process of image captioning is shown in Fig. 1.

The original image captioning algorithm is mainly realized by using templates. This method first extracts a series of key feature data such as key objects and special attributes through different types of classifiers, such as SVM, and then uses lexical model or other specific templates to convert the obtained feature information into description text. The key feature information extracted by Farhadi et al. [3] mainly includes three information: the object in the image, the action of the object and the scene in which the object is located. The noise item is removed by some common smoothing methods. Finally, the relationship between the extracted information is analyzed and the final image description result is obtained; Li et al. [4] mainly extracted the meaningful phrase information in the image, and then combined the phrases by n-gram dynamic fusion to select the best phrase combination to generate image description; The feature information extracted by Kulkarni et al. [5] also includes the object in the image. In addition, it also uses the detector to directly extract the attributes of the object and the relationship between the attribute and the object, and generates a triple information containing the extracted information by using CRF, and finally generates the description by using the established rules. The above methods rely too much on fixed rules and extracted specific information, and different rules will directly affect the generated statements. The image description content obtained by the above methods is too simple and lacks versatility, which often fails to achieve the desired effect.

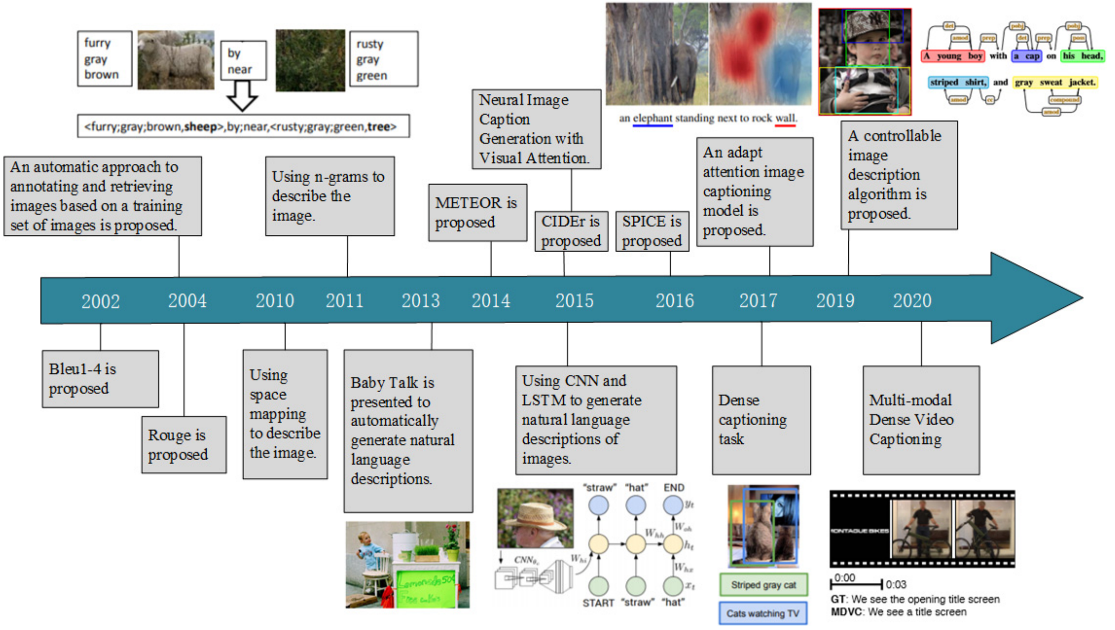

furry
gray
brown
by
near
rusty
gray
green
<furry;gray;brown,sheep>,by;near,<rusty;gray;green,tree>
Neural Image Caption Generation with Visual Attention.
an elephant standing next to rock wall.
A young boy with a cap on his head, striped shirt, and gray sweat jacket.
det
amod
prep
pobj
poss
cc
compound
An automatic approach to annotating and retrieving images based on a training set of images is proposed.
Using n-grams to describe the image.
METEOR is proposed
CIDEr is proposed
SPICE is proposed
An adapt attention image captioning model is proposed.
A controllable image description algorithm is proposed.
2002
2004
2010
2011
2013
2014
2015
2016
2017
2019
2020
Bleu1-4 is proposed
Rouge is proposed
Using space mapping to describe the image.
Baby Talk is presented to automatically generate natural language descriptions.
Using CNN and LSTM to generate natural language descriptions of images.
Dense captioning task
Multi-modal Dense Video Captioning
"straw"
"hat"
END
START
y_t
h_t
x_t
W_{oh}
W_{hh}
W_{hx}
W_{hi}
CNN_{θ_c}
Striped gray cat
Cats watching TV
0:00
0:03
GT: We see the opening title screen
MDVC: We see a title screen

Figure.1 Development history.

The image captioning model based on encoder decoder structure is more flexible in prediction. In the encoder, convolution neural network is used to extract and vectorize the main image feature information; The decoder mainly uses recurrent neural network, which combines the vectorized image features with semantic information to generate a description statement of image content.

The model in reference [6-7] is simple, only the image features are transferred in the first step of decoding. As the time step of the prediction sequence is longer, the important image feature information is lost, so that the effect of image prediction is not ideal. With the emergence of attention mechanism, researchers began to use attention mechanism to solve the problem of image captioning. Xu et al. [8]

proposed two attention mechanisms and applied them to different types of image regions. The soft attention mechanism focuses on all regions of the image, and the weight value of each region is different. The higher the value is, the more important the region is in prediction. Finally, the context vector to be used is obtained by the weight distribution of each region. Unlike the soft attention mechanism, the hard attention mechanism only focuses on a certain area, and in order to increase flexibility, it is usually adopted Select an area randomly and calculate the probability value of the selected area. You et al. [9] proposed semantic attention, and the model structure is shown in Fig. 2. The model combines the bottom-up and top-down image description modes, The network can not only judge how to use attention flexibly and accurately according to the important features of images, but also take the important information in semantics as the focus of attention. Firstly, the image feature information processed by convolution neural network is used to make the recurrent neural network contain the visual feature information; then, the attrdet attribute detector is used to detect the visual objects or attributes; finally, the two are fused to obtain the final image description sequence by using the recurrent neural network.

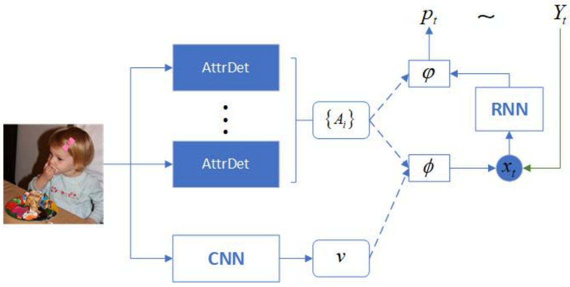


Figure.2 Semantic attention model structure.

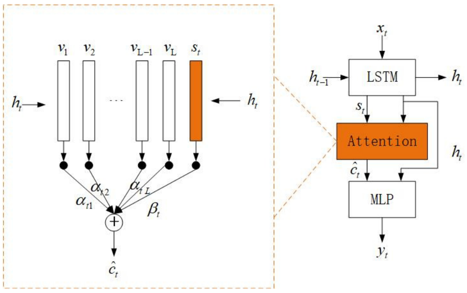


Figure.3 Adaptive attention model structure.

Lu et al. [10] proposed an adaptive attention mechanism. The model structure is shown in Fig. 3. The reason for this model is that every time people describe the content of an image, almost only two kinds of words are included in the image features, one can be called visual words, which can be well predicted by using image information; the other is called non visual words, which are words in the form of "and", "to" and "a", which can not be obtained by analyzing image information, It can only be inferred from the semantic information of the context. The model adds a visual guidance sign containing a certain amount of semantic information, which is used to determine whether attention mechanism is needed according to the situation. If the mechanism is needed, the multi region features obtained by image convolution will be assigned different proportions.

Chen et al. [11] proposed spatial and channel level attention. The model structure is shown in Fig. 4. In this model, spatial attention and channel level attention are fused, and the combined information is used to generate image description. This model is no longer like the previous attention mechanism, it can use the context semantic information of the collected text sequence at the encoding end. For example, when we want to predict kites, network channel level attention will pay more attention to the semantic information of "Kite", "sky" and "people", so as to get channel level feature map.Yao et al. [12] proposed a way to use dual image attribute features. It not only uses the image features obtained by convolution neural network, but also uses the high-level attribute information of the image. The high-level attribute information is mainly extracted by multi instance learning method, which makes the image features more abundant and makes it easy to obtain more accurate image description statements.

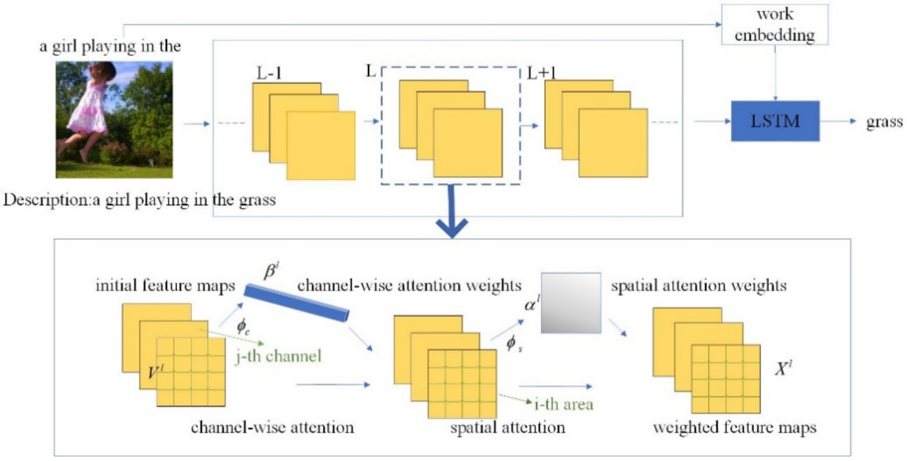


Figure.4 Structure of attention model at spatial and channel levels.

In recent years, there are many image captioning models based on encoder decoder architecture, For example, Cornia et al. [13] proposed an image captioning model that can be controlled by signals and a multimodal image captioning model combining visual and auditory information proposed by Iashin et al. [14] , these models have good prediction effect.

3. Datasets and Evaluation Indexes

This chapter mainly introduces the common datasets and various evaluation indexes of image captioning algorithm. In order to achieve a good network model, we must have data sets as support. By training a large number of data sets, we can not only avoid the occurrence of over fitting phenomenon, but also make the network model more perfect. In order to judge the quality of a model, we must use the evaluation index. We can see the advantages and disadvantages of the model through the numerical value of each evaluation index.

3.1. Datasets

Image captioning algorithms need to contain both image and image corresponding description type data sets. The data sets related to this type mainly include Visual Genome [15] , ICC [16] , flickr8k [17] , flickr30k [18] and MSCOCO 2014 [19] . Visual genome was developed by Michael Bernstein collects a data set of image intensive annotation, which contains 108,077 image data. Each image contains about 35 objects, and each object will be labeled. Each image contains 50 regional descriptions, with 540,000 domain descriptions in total. In addition, there are 170,000 QA pairs and 280,000 image attributes There are 230,000 image relations. ICC is a dataset that describes images in Chinese and is a part of AIC dataset. It mainly contains 300,000 images. Each image contains five corresponding Chinese description statements, totaling 1,500,000 statements. Flickr8k and flickr30k are data sets collected by flickr that describe various human activities. Flickr8k mainly contains 8,000 images, 6,000 of which are used for training, the remaining 1,000 for verification and 1,000 for testing. Flickr30k mainly contains 31,000 images, including 29,000 training sets, 1,000 verification sets and test sets. MSCOCO 2014 is an image data set established by Microsoft that can be used for object detection, semantic segmentation and image captioning. The data set has been published many times. Mscoco 2014 contains 82,783 training samples, 40,504 verification samples and 40,775 test samples. In addition, there are 270,000 segmented portrait images and 886,000 segmented object images. Compared with other data sets, the MSCOCO 2014 dataset has more abundant examples for image captioning annotation, so it is widely used in image captioning research.

3.2. Evaluation Indexes

The evaluation indexes of image captioning are mainly divided into BLEU1-4 [20], ROUGE [21], METEOR [22], SPICE [23] and CIDEr [24]. BLEU was originally used to evaluate the quality of translation results. It mainly includes three parts: precision value of n-gram coincidence ratio; penalty mechanism, mainly to solve the problem of too small length of prediction sequence; geometric average, mainly to balance the difference of precision values of n-gram. ROUGE is a method designed to evaluate the effect of the generated text summary. ROUGE mainly includes four different evaluation index, namely ROUGE _ N、ROUGE_ L、ROUGE_ W and ROUGE _ S. The latter three methods can be applied to the evaluation of image captioning prediction effect. METEOR is also an evaluation index used to evaluate the translation effect. METEOR uses a new punishment mechanism, which is mainly used to punish the case that the predicted sequence does not correspond to the real tag word order. At the same time, METEOR uses chunks to replace the unigram used in Bleu, and each chunk is composed of adjacent unigram words. SPICE avoids the overlapping problem caused by using n-gram. Firstly, it encodes the predicted sequence and the real tag into a semantic dependency tree by using dependency parser, and maps it. After mapping, the map can be divided into tuples containing objects, relationships and attributes. Finally, it matches the graph obtained from the predicted sequence with the map from the real label to get the evaluation results. CIDEr does not judge the quality of prediction description by sentence matching, but evaluates the quality by the similarity between the predicted text and the real label. The higher the similarity between the two, the more suitable the predicted description is and the better the prediction effect is. In recent years, compared with the previous several evaluation indicators, this index is used most frequently in evaluating image description quality.

4. Conclusion

This paper introduces some existing image captioning methods and analyzes their principles. The data sets and evaluation indexes needed in this field are introduced. Although the existing image captioning algorithms have improved the prediction effect to a certain extent, they do not realize the function of generating specific description statements according to specific situations.

In the future, image captioning can be improved in three aspects. The first aspect is to enhance the adaptability of the network, so that the model can pay attention to the specific situations and concerns, and generate targeted descriptions according to different situations and concerns. The second aspect is to optimize the evaluation algorithm to evaluate the quality of the output sequence of the model more accurately. The third aspect is to enhance the robustness of the model and avoid the influence of interference characteristics on the model output.

References

[1] Lee, S., & Kim, I. (2018). Multimodal feature learning for video captioning. Mathematical Problems in Engineering, 2018.

[2] Kim, H., Tang, Z., & Bansal, M. (2020). Dense-Caption Matching and Frame-Selection Gating for Temporal Localization in VideoQA. arXiv preprint arXiv:2005.06409.

[3] Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., & Forsyth, D. (2010). Every picture tells a story: Generating sentences from images. In: European conference on computer vision, Berlin, pp. 15-29.

[4] Li, S., Kulkarni, G., Berg, T., Berg, A., & Choi, Y. (2011). Composing simple image descriptions using web-scale n-grams. In: Proceedings of the Fifteenth Conference on Computational Natural Language Learning, pp. 220-228.

[5] Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., & Berg, T. L. (2013). Babytalk: Understanding and generating simple image descriptions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(12): 2891-2903.

[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] Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3156-3164.

[8] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning, pp. 2048-2057.

[9] You, Q., Jin, H., Wang, Z., Fang, C., & Luo, J. (2016). Image captioning with semantic attention. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4651-4659.

[10] Lu, J., Xiong, C., Parikh, D., & Socher, R. (2017). Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 375-383.

[11] Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., & Chua, T. S. (2017). Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5659-5667.

[12] Yao, T., Pan, Y., Li, Y., Qiu, Z., & Mei, T. (2017). Boosting image captioning with attributes. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4894-4902.

[13] Cornia, M., Baraldi, L., & Cucchiara, R. (2019). Show, control and tell: A framework for generating controllable and grounded captions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8307-8316.

[14] Iashin, V., & Rahtu, E. (2020). Multi-modal Dense Video Captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 958-959.

[15] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., ... & Bernstein, M. S. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision, 123(1): 32-73.

[16] Wu, J., Zheng, H., Zhao, B., Li, Y., Yan, B., Liang, R., & Wang, Y. (2017). Ai challenger: A large-scale dataset for going deeper in image captioning. arXiv preprint arXiv:1711.06475.

[17] Hodosh, M., Young, P., & Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research, 47: 853-899.

[18] Young, P., Lai, A., Hodosh, M., & Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics, 2: 67-78.

[19] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In: European conference on computer vision, Cham, pp. 740-755.

[20] 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.

[21] Lin, C. Y. (2004). Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out, pp. 74-81.

[22] Denkowski, M., & Lavie, A. (2014). Meteor universal: Language specific translation evaluation for any target language. In: Proceedings of the ninth workshop on statistical machine translation, pp. 376-380.

[23] Anderson, P., Fernando, B., Johnson, M., & Gould, S. (2016). Spice: Semantic propositional image caption evaluation. In: European Conference on Computer Vision, Cham, pp. 382-398.

[24] Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). Cider: Consensus-based image

description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4566-4575.