Artificial Intelligence and Social Computing, Vol. 72, 2023, 58–66
https://doi.org/10.54941/ahfe1003274

Image Caption Generation of Arts: Review and Outlook

Baoying Zheng and Fang Liu

Hunan University, Changsha, Hunan, 410082, China

ABSTRACT

Image captioning extract image features and automatically describe the content of an image in words. Recently image captioning has broken through the application of natural images and is widely used in the arts. It can be applied to art retrieval and management, and it can also automatically provide artistic introductions for the visually impaired. This paper reviews related research in image captioning of artworks, and divides image captioning into three types, including template-based approach, retrieval-based approach, and generative approach. Furthermore, mainstream generative approaches include Encoder-decoder, Transformer, New generation framework, etc. Finally, this paper summarizes the evaluation metrics for image captioning, and looks forward to the application and future development of art image captioning.

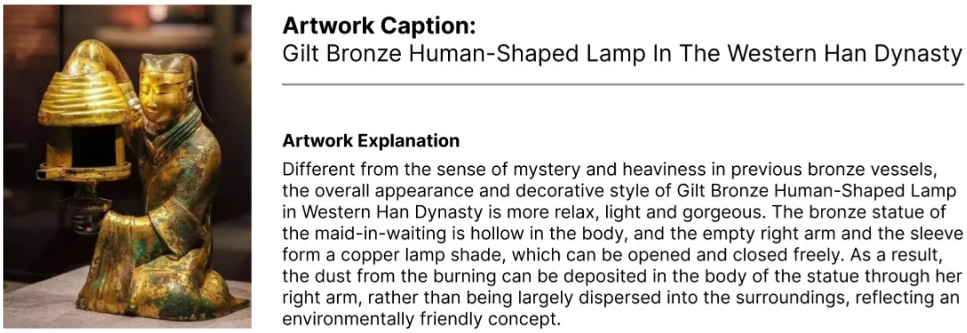
Keywords: Image captioning, Artwork analysis, Deep neural network

INTRODUCTION

Image captioning extracts relevant features from an image and predicts one or more sequences of textual descriptions, which requires the combination of computer vision and natural language processing methods. The goal is to build a model that is able to understand the content of an input image and to generate a natural language description as output (Allaouzi et al. 2018). Recently, image captioning task in the arts has received growing attention and a large amount of interest from deep learning, computer vision, and natural language processing researchers.

Such a captioning system could be used for automatically generating explanations of artworks (see Figure 1), and it allows a visitor of a museum or cultural heritage site to obtain a detailed description of artwork on their mobile device by just taking a picture of the artwork (Sheng and Moens, 2019). It facilitates a deeper interaction between the general public and artworks. Besides, image captioning assists in retrieval from online digital archives, as well as hereby saving time and labor of manual annotation that potentially ease the work of art curators by automatically generating comments of artworks (Bai et al. 2021).

However, image captioning is a non-trivial task due to various challenges that are unique to the arts, including the lack of large-scale training data of image-text pairs, the complexity of meaning associated with artworks, and the need for expert-level annotations, there are still many limitations in



Artwork Caption:
Gilt Bronze Human-Shaped Lamp In The Western Han Dynasty

Artwork Explanation
Different from the sense of mystery and heaviness in previous bronze vessels, the overall appearance and decorative style of Gilt Bronze Human-Shaped Lamp in Western Han Dynasty is more relax, light and gorgeous. The bronze statue of the maid-in-waiting is hollow in the body, and the empty right arm and the sleeve form a copper lamp shade, which can be opened and closed freely. As a result, the dust from the burning can be deposited in the body of the statue through her right arm, rather than being largely dispersed into the surroundings, reflecting an environmentally friendly concept.

Figure 1: An example of artwork image, caption and explanation from specialized museum.

image captioning in the arts. Art image captioning approaches can be divided into three broad categories: (1) template-based approaches that fill the detected objects into the blanks in the predefined sentences. (2) retrieval-based approaches that involve retrieving the most suitable caption from a database of image-caption pairs and assigning it to a novel image. (3) generative approaches that generate novel captions using deep neural networks.

This paper mainly introduces the second and third models and summarizes evaluation metrics in the next. New image captioning methods for artworks explanation will be explored in the last.

Image Captioning Approaches Categories

Several models have been proposed to address the task of image captioning. We can broadly classify them into the following three categories:

Template-based Approach

This approach consists of three parts: the predefined sentences with blanks in it, the object detection model, and the relation model. The blanks in the predefined sentences are properly filled with detected objects, such as entities, attributes, and behaviors (Hutchison et al. 2010) (Li et al. 2011) (Kulkarni et al. 2013).

Lei and Wang (2015) and Xu et al. (2017) introduce respectively an ontology and hierarchical model to describe the creation of Dunhuang frescoes. The two works leverage low-level image features, including the image texture and meta-data of cultural images but encode them in different-structured models. Nevertheless, the captions generated by this approach are more likely to be grammatically correct, its main drawback is that it relies on hard-coded visual concepts, which constrained the flexibility and the variety of the output.

Retrieval-Based Approach

This approach casts the task as a retrieval problem (Hossain et al. 2019). the idea is to search the image most similar to the input image in the large dataset and then either the caption is directly transferred to the novel image (Farhadi et al. 2010) (Hodosh et al. 2013) (Socher et al. 2014), or alternatively a new caption is generated by combining fragments of the candidate captions according to certain rules or schemes (Kuznetsova et al. 2014) (Li et al. 2011).

Garcia and Vogiatzis (2018) introduce the SemArt dataset, a collection of fine-art images associated with textual comments, intending to map the images and their descriptions in a joint semantic space. They compare different combinations of visual and textual encodings, in projecting the visual and textual encodings in a common multimodal space. Baraldi et al. (2018) address the task of creating a shared embedding space to retrieve the associations between images and texts for artworks. The authors introduce a new visual semantic dataset named BibleVSA, a collection of miniature illustrations and commentary text pairs, and explore supervised and semi-supervised approaches to learning cross-references between textual and visual information.

Wang et al. (2006) propose an ALIP (Automatic Linguistic Indexing of Pictures) system of pictures to built to learn the expertise of a human annotator based on a small-scale annotated EMPEROR Collection. The authors first train a 2-D multiresolution hidden Markov model (2-D MHMM) to find the correspondence between a cultural image and its descriptive concepts. When an un-annotated image is presented, the system computes the statistical likelihood of the image resembling each of the learned statistical models, and the best concept is selected to annotate the image. However, these models are based on the assumption that similar images have the same captions, and they cannot generate image-specific captions.

Deep Neural Networks Generative Approach

This approach exploits the use of deep neural networks, which have led to huge success in the field of computer vision and natural language generation and also in the interplay between them (Allaouzi et al. 2018).

Encoder-decoder A limited number of studies contributed to the task of generating descriptions of artwork images using deep neural networks and most of them rely on employing the encoder-decoder architecture-based image captioning approach (Cetinic, 2021). This models uses an end-to-end trainable network to learn the mapping from images to captions directly. and it employs encoder-decoder architecture by combining a Convolutional Neural Network (CNN) with a Recurrent Neural Network (RNN). Where, CNN is used as an image encoder to produce a fixed-length vector representation, which is then fed into the decoder RNN to generate a caption (Donahue et al. 2015) (Vinyals et al. 2015).

Sheng and Moens (2019) proposes an encoder-decoder framework for generating captions of artwork images where the encoder (ResNet18 model) extracts the input image feature representation and the artwork type representation, while the decoder is a long short-term memory (LSTM) network to generate a corresponding caption based on the input image vector. They introduce two image captioning datasets referring to ancient Egyptian and Chinese art. However, the generated captions are short, do not support the user's understanding well, and often produce irrelevant content with low accuracy and intelligibility.

Another very recent work Gupta et al. (2020) present a novel captioning dataset for art historical images consisting of 4000 images across 9 iconographies, along with a description for each image consisting of one or more paragraphs. They used this dataset to fine-tune different variations of image captioning models based on the encoder-decoder approach introduced in Vinyals et al. (2015). Kuang-Yu et al. (2017) select a particular aesthetic aspect and generate captions with respect to the aspect chosen. The approach is proposed by training the CNN model with a soft-attention layer to predict the aspect-fusion coefficients from the context information, which could leverage the hidden annotations of different aspects and choose the proper combination dynamically over time to generate a semantically meaningful caption.

Transformer Cetinic (2021) uses the artwork image dataset to fine-tune a transformer-based vision-language pre-trained model. The available annotations are processed into clean textual descriptions, and the existing dataset is transformed into a collection of suitable image-text pairs. Object classification aware region features are extracted from the images using the Faster RCNN model and employ a unified vision-language pre-training model (VLP) in fine-tuning experiment.

New Generation Framework Bai et al. (2021) introduce a multi-topic and knowledgeable art description framework, which modules the generated sentences according to three artistic topics and, additionally, enhances each description with external knowledge. The authors extract D-dimensional visual features from L spatial locations from the image (Xu et al. 2015) using a pre-trained ResNet (He et al. 2016), $V = \{\mathrm{v}_1, \mathrm{v}_2, ..., \mathrm{v}_L\}$, $\mathrm{v}_\mathrm{i} \in \mathrm{R}^D$. Then, in the masked sentence generation part, they input V into the topic decoder to generate multi-topic masked sentences that describe the painting from multiple aspects. In the knowledge retrieval part, they take the average-pooling vector $\mathrm{v} = \Sigma_\mathrm{i}\ \mathrm{v}_\mathrm{i}/L$ as the global visual feature for multi-attribute prediction. The predicted attributes and the detected objects are used to retrieve relevant knowledge from an external source with DrQA (Chen et al. 2017). Finally, given the generated multi-topic masked sentences and the retrieved knowledge text, they extract candidate knowledge concepts, and use a BERT-based model (Devlin et al. 2019) to get the final description.

Jin et al. (2019) propose Aesthetic Multi-Attribute Network (AMAN) to produce captions for each image aesthetic attribute, which contains a multi-attribute feature network (MAFN), channel and spatial attention network (CSAN), and language generation network (LGN). MAFN measures the feature matrix of 5 attribute scores through the multi-task regression. Multi-attribute networks are pre-trained on DPC-Captions dataset and fine-tuned on their weakly-annotated large-scale dataset. The CSAN dynamically adjusts the attentional weights of channel dimension and spatial dimension of the obtained features. Finally, LGN generates the captions by LSTM network which needs ground truth attribute captions in DPC-Captions and adjusted feature maps from CSAN.

New interactive captioning method Besides, image captioning can assist the visually impaired in accessing and appreciating visual art. Ahmetovic et al. (2021) introduce novel interaction techniques for artwork captions exploration. The techniques leverage touch screen as the interface for scanning the artwork image area, while additional information on the explored elements is provided through verbal feedback. Caption segmentation is designed by domain experts, and it presents a hierarchical segmentation of the artwork elements and more verbose descriptions. Although they do not automatically generate captions, they provide a new artistic caption interaction.

Evaluation Metrics

Image captioning metrics include the automatic manner and human evaluation. Some of the earliest attempts in automatic manners have originated from the readily available metrics for machine translation, including the standard BLEU (Papineni et al. 2001), METEOR (Denkowski and Lavie, 2014), ROUGE (Lin, 2004). BLEU (Papineni et al. 2001) is defined as the geometric mean of n-gram (1 to 4-gram) precision scores multiplied by a brevity penalty for short sentences. ROUGE (Lin, 2004) measures the recall of n-grams and therefore rewards long sentences. Specifically, ROUGE-L measures the longest matching sequence of words between a pair of sentences. METEOR (Denkowski and Lavie, 2014) represents the harmonic mean of precision and recall of unigram matches between sentences and additionally includes synonyms and paraphrase matching.

CIDEr (Vedantam et al. 2015) and SPICE (Anderson et al. 2016) metrics are specifically developed for image caption evaluation tasks. CIDEr (Vedantam et al. 2015): measures the average cosine similarity between the candidate sentence and the reference sentences by performing a Term Frequency Inverse Document Frequency (TF-IDF) weighting for each n-gram. SPICE (Anderson et al. 2016): measures the quality of generated captions by computing an F-score based on the propositional semantic content of candidate and reference sentences represented as scene graphs.

However, it remains questionable how adequate these metrics are in assessing the overall quality of the captions in this particular art context. All of the reported metrics mostly measure the word overlap between generated and reference captions. They are not designed to capture the semantic meaning of a sentence and often lead to poor correlation with human judgment. The generated caption could be semantically aligned with the image content but represent a different version of the original caption and therefore have very low metric scores. Therefore a human qualitative analysis of the results is also required in order to better understand potential contributions and drawbacks of the proposed approach.

Outlook for Future Research Work

Standard image captioning mainly focuses on natural images and only the content is considered. However, For the general public, art tends to be considered as a mysterious and remote discipline that requires a lot of studies to be fully appreciated (Bai et al. 2021). So we need to address what kind of description would be regarded as “adequate” for art particular purpose. Considering for instance Erwin Panofsky’s three levels of analysis (Panofsky, 1991), we can distinguish the “pre-iconographic” description, “iconographic” description and the “iconologic” interpretation as possibilities of aligning semantically meaningful. While captions of natural images usually function on the level of “pre-iconographic” descriptions, which implies simply listing the elements that are depicted in an image, for artwork images this type of description represents only the most basic level of visual understanding and is often not considered to be of great interest.

So a comprehensive explanation of an artwork requires not only the factual description of its content, but also background knowledge, such as details about its author, the context of the creation process, cultural symbols and so on(see Figure 2). But this information is rarely contained in the artwork image itself (Sheng and Moens, 2019). Generally speaking, “iconologic” interpretation have great potential in mining the emotions, stories, connotations, and metaphors of artworks.

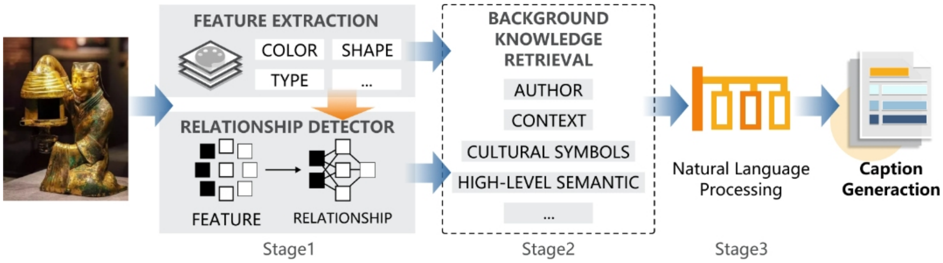


Figure 2: A new framework of artwork image caption generation, it contains features extraction, relationship detector, external knowledge retrieval and natural language processing.

CONCLUSION

This paper reviews the approaches of image captioning in the arts. The mainstream approach is to use Deep Neural Networks to retrieve the most similar image’s description or generate novel descriptions, including Encoder-decoder, Transformers and New Generation Frameworks designed for specific tasks. Auto-generated image captioning can be used in art management and combined with new interactive methods to provide cultural services to the visually impaired. Finally, we discuss future research trends for image captioning, which can generate high-level semantic descriptions beyond the content of artistic images that explain the cultural connotation of artworks understandably.

ACKNOWLEDGMENT

This work is supported by The National Key Research and Development Program of China [2019YFB1405702].

REFERENCES

Ahmetovic, D., Kwon, N., Oh, U., Bernareggi, C. and Mascetti, S. (2021). Touch Screen Exploration of Visual Artwork for Blind People. In: Proceedings of the Web Conference 2021. Ljubljana Slovenia: ACM, pp. 2781–2791.

Allaouzi, I., Ben Ahmed, M., Benamrou, B. and Ouardouz, M. (2018). Automatic Caption Generation for Medical Images. In: Proceedings of the 3rd International Conference on Smart City Applications. Tetouan Morocco: ACM, pp. 1–6.

Anderson, P., Fernando, B., Johnson, M., & Gould, S. (2016). Spice: Semantic propositional image caption evaluation. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14, pp. 382–398. Springer International Publishing.

Bai, Z., Nakashima, Y. and Garcia, N. (2021). Explain Me the Painting: Multi-Topic Knowledgeable Art Description Generation. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada: IEEE, pp. 5402–5412.

Baraldi, L., Cornia, M., Grana, C., & Cucchiara, R. (2018). Aligning text and document illustrations: towards visually explainable digital humanities. In 2018 24th International Conference on Pattern Recognition (ICPR), pp. 1097–1102. IEEE.

Cetinic, E. (2021). Iconographic image captioning for artworks. In: Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part III, pp. 502–516. Springer International Publishing.

Chen, D., Fisch, A., Weston, J., & Bordes, A. (2017). Reading wikipedia to answer open-domain questions. arXiv preprint arXiv:1704.00051.

Denkowski, M. and Lavie, A. (2014). Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In: Proceedings of the Ninth Workshop on Statistical Machine Translation. Baltimore, Maryland, USA: Association for Computational Linguistics, pp. 376–380.

Devlin, J., Chang, M. W., & Lee, K. (2019). Google, KT, Language, AI: BERT: pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT, pp. 4171–4186.

Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Darrell, T. and Saenko, K. (2015). Long-term recurrent convolutional networks for visual recognition and description. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: IEEE, pp. 2625–2634.

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 Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11, pp. 15–29. Springer Berlin Heidelberg.

Garcia, N., & Vogiatzis, G. (2018). How to read paintings: semantic art understanding with multi-modal retrieval. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, pp. 0–0.

Gupta, J., Madhu, P., Kosti, R., Bell, P., Maier, A., & Christlein, V. (2020). Towards image caption generation for art historical data. In: Proceedings of the AI Methods for Digital Heritage, Workshop at KI2020 43rd German Conference on Artificial Intelligence, Bamberg, Germany, pp. 21–25.

He, K., Zhang, X., Ren, S. and Sun, J. (2016). Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, pp. 770–778.

Hodosh, M., Young, P. and Hockenmaier, J. (2013). Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. Journal of Artificial Intelligence Research [online], 47, pp. 853–899.

Hossain, MD. Z., Sohel, F., Shiratuddin, M. F. and Laga, H. (2019). A Comprehensive Survey of Deep Learning for Image Captioning. ACM Computing Surveys [online], 51(6), pp. 1–36.

Hutchison, D. et al. (2010). Every Picture Tells a Story: Generating Sentences from Images. In: Daniilidis, K., Maragos, P., and Paragios, N., eds. Computer Vision – ECCV 2010. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 15–29.

Jin, X., Wu, L., Zhao, G., Li, X., Zhang, X., Ge, S., Zou, D., Zhou, B. and Zhou, X. (2019). Aesthetic Attributes Assessment of Images. In: Proceedings of the 27th ACM International Conference on Multimedia. Nice France: ACM, pp. 311–319.

Kuang-Yu Chang, Lu, K.-H. and Chen, C.-S. (2017). Aesthetic Critiques Generation for Photos. In: 2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, pp. 3534–3543.

Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., A. C. Berg, T. L. (2013). Babytalk: Understanding and generating simple image descriptions. IEEE transactions on pattern analysis and machine intelligence, 35(12), 2891–2903.

Kuznetsova, P., Ordonez, V., Berg, T. L. and Choi, Y. (2014). T ree T alk : Composition and Compression of Trees for Image Descriptions. Transactions of the Association for Computational Linguistics [online], 2, pp. 351–362.

Lei, X. and Wang, X. (2015). Semantic Description of Cultural Digital Images: Using a Hierarchical Model and Controlled Vocabulary. D-Lib Magazine [online], 21(5/6).

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.

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

Panofsky, E. (1991). Studies in iconology: humanistic themes in the art of the Renaissance. 14. [pr.]. New York: Harper & Row.

Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J. (2001). BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL'02. Philadelphia, Pennsylvania: Association for Computational Linguistics, p. 311.

Sheng, S. and Moens, M.-F. (2019). Generating Captions for Images of Ancient Artworks. In: Proceedings of the 27th ACM International Conference on Multimedia. Nice France: ACM, pp. 2478–2486.

Socher, R., Karpathy, A., Le, Q. V., Manning, C. D. and Ng, A. Y. (2014). Grounded Compositional Semantics for Finding and Describing Images with Sentences. Transactions of the Association for Computational Linguistics [online], 2, pp. 207–218.

Vedantam, R., Zitnick, C. L. and Parikh, D. (2015). CIDEr: Consensus-based image description evaluation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: IEEE, pp. 4566–4575.

Vinyals, O., Toshev, A., Bengio, S. and Erhan, D. (2015). Show and tell: A neural image caption generator. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: IEEE, pp. 3156–3164.

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.

Wang, J. Z., Grieb, K., Zhang, Y., Chen, C., Chen, Y. and Li, J. (2006). Machine annotation and retrieval for digital imagery of historical materials. International Journal on Digital Libraries [online], 6(1), pp. 18–29.

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

Xu, L., Merono-Penuela, A., Huang, Z., & Van Harmelen, F. (2017). An ontology model for narrative image annotation in the field of cultural heritage. In: Proceedings of the 2nd Workshop on Humanities in the Semantic web (WHiSe), pp. 15–26.