A Survey on Recent Advances in Image Captioning

Feng Chen¹, Xinyi Li²,³, Jintao Tang¹, Shasha Li¹ and Ting Wang¹

¹ Department of Computing Science, College of Computer, National University of Defense Technology, China
² Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, China
³Corresponding author’s Email: lixinyimichael@163.com

Abstract. Image captioning, an interdisciplinary research field of computer vision and natural language processing, has attracted extensive attention. Image captioning aims to produce reasonable and accurate natural language sentences to describe images. It requires the captioning model to recognize objects and describe their relationships accurately. Intuitively, it is difficult for a machine to have the general image understanding ability like human beings. However, deep learning provides the basis for intelligent exploration. In this review, we will focus on recent advanced deep methods for image captioning. We classify existing methods into different categories and discuss these categories respectively; meanwhile, we discuss the related datasets and evaluation metrics. We also prospect the future research directions.

1. Introduction
With the development of intelligent terminal and social network technology, multi-modal data (e.g., image, text, video) has increased dramatically, such as posts and articles on the microblog, online news, and advertisements. Massive data is the basis of the development of modern intelligent technology, especially the data covering multiple modals. If the information in different modals can be aligned, it will undoubtedly power machine intelligence and help solve practical social problems. For human beings, it is easy to transform between modals, because life experiences and continuous practices bring us the ability of multi-modal understanding. However, multi-modal transformation remains a challenging task for machines.

This paper focuses on the transformation from image to natural language sentence: image captioning. Compared with traditional computer vision tasks such as object tracking [1] and multi-modal retrieval [2], image captioning is more challenging. It requires the model to further weigh the scene of each object and relationships between objects, to generate correct natural language descriptions, syntactically and semantically.

As an essential step to explore machine intelligence, image captioning has wide application scenarios. For example, for people with visual impairment, such as blindness and color blindness, image captioning can help them achieve visual understanding; in the image retrieval task, image captioning helps the model achieve more accurate content-based retrieval; in the field of intelligent robots, image captioning can be embedded into intelligent systems to help machines gain better visual understanding.

Thanks to the successful application and rapid development of deep learning technology, image captioning has achieved remarkable success in the past few years.
There are reviews in this field that well summarize the work in this field. However, the recent advances are not fully covered. This paper is a review of the latest progress in this field to fill this gap. We only focus on the recent advances and divide these methods into different categories.

2. Recent advances for image captioning

2.1. Methods with the encoder-decoder framework

The encoder-decoder framework is the basis of most deep methods, shown in Figure 1. In a typical method of this subcategory, a CNN (convolution neural network) is utilized as the encoder to encode the input image into a representation, and then an LSTM-based language model is used as the decoder to produce a caption word by word with the image representation. Vinyals et al. [3] first proposed a NIC model with the encoder-decoder framework for image captioning. In this architecture, the current generated word is based on the previous hidden state of the LSTM and the prior word. However, the image representation is only fed into the LSTM at the initial step and its influence on the posterior word becomes weaker. Therefore, Jia et al. [4] proposed a novel LSTM model, which worked at a global scale. By applying global semantic information of images as the extra guidance of LSTM cells, this work managed to generate more high-quality captions. Since previous methods [3, 4] are based on unidirectional and shallow LSTMs, to make full use of both the history and future context, Wang et al. [5] designed an end-to-end bidirectional LSTM model. This model used two LSTMs to exploit both long-term back and forward contexts. Meanwhile, the proposed model explored its deeper variants.

2.2. Attention-enhanced methods

To capture more fine-grained visual information during generation, attentive methods are proposed to ground generated words on salient image parts. Xu et al. [6] is the first to introduce attention mechanisms for image captioning: a hard attention mechanism and a soft attention mechanism. This method focuses on the most relevant image regions during generation, yet a drawback stands out: it utilizes image features from the lower CNN layer which may fail to capture high-level information. Jin et al. [7] extracted visual information of particular objects, and predicted topic vectors of images which were then used as the contexts during generation. Rather than associating image features only to the decoder state, Pedersoli et al. [8] proposed the area-based attention mechanisms. It models associations between the generated words and the image regions. Lu et al. [9] considered that not all of the caption words have corresponding visual parts thus they introduced a novel adaptive attention mechanism that can decide when to rely on image regions and when to only rely on the previously generated words of the language model. Anderson et al. [10] proposed to combine both the Bottom-Up and the Top-Down attention mechanisms for caption generation: Bottom-Up proposed a set of image regions represented by pooled convolutional feature vectors and Top-Down calculated attention weights over these regions during generation.

A traditional attention-based model aligns the attended image feature to one captioning word at each time step, therefore, Huang et al. [11] further proposed a model to align image features and the target words adaptively for caption generation. In this method, a base attention model takes one single attention step per time step, then an adaptive attention module learns how many times to attend. The output of a previous attention mechanism directly depends on the current attention result, which lacks relationship modeling between the attended feature and the attention query. To tackle this problem, Huang et al. [12] further presented an attention-on-attention (AoA) module to model the relevance between the attention result and the attention query. Based on the attention result and the query, it
firstly generates an information vector and an attention gate, then utilizes another attention operation to them, thus the finally the expected knowledge is obtained.

2.3. Semantic-enhanced methods
Semantic-enhanced methods exploit the high-level semantic features of images for captioning, e.g., attributes, concepts, and structural semantics. Wu et al. [13] first introduced high-level semantics into the encoder-decoder framework and verified the effectiveness of explicit representations of attributes for image captioning. Yao et al. [14] further presented a novel LSTM with attributes, injecting attribute information into the captioning model. They explore the inter-attribute relationships by multiple instance learning and augment attributes to complement image representations. Zhang et al. [15] proposed to use part of speech guidance. They utilize attention-based methods to take advantage of the generated part of speech information and use the multi-task learning strategy to facilitate training. Huang et al. [16] modeled the similarity between an attribute concept and an object, and selected subsequent attributes to attend to during generation. In this method, one generated word is relevant to a sub-sequence of attributes, enhancing semantic explainability.

Most recently, structural semantics have been explored. Chen et al. [17] used the Abstract Scene Graph (ASG) to decide what and how detailed to say in the final description. ASG is a directed graph that contains objects, attributes, and relations grounded in visual features, but it has nothing to do with explicit semantics. Yang et al. [18] proposed a scene graph auto-encoder to exploit the language inductive bias as a priori. They use a scene graph to represent structural semantics information of an image and learn a dictionary to reconstruct the caption.

2.4. Transformer-based methods
Image captioning models based on transformer [19] has achieved great success. Li et al. [20] investigated a transformer-based model with only the attention and the feed-forward layers, which exploited both the visual and the semantic information. Technically, they apply a bilateral gating mechanism to control the propagation of the visual and the semantic information. To find a way to capture higher-order interactions, Pan et al. [21] designed x-linear attention blocks based on bilinear pooling, in which the spatial- and the channel-wise bilinear attention features capture higher-order interactions among the input features. Cornia et al. [22] proposed a meshed-memory transformer for image captioning which modeled the relationships among regions and learned a priori knowledge, and designed a meshed connection for decoding. Since previous transformer-based methods only focus on region-level information without considering the global representation of an image, Ji et al. [23] Leveraged the intra- and inter-layer global representation in transformer for captioning. It abstracts a comprehensive global representation of the input image as guidance to encourage the decoder to generate better captions.

2.5. Post-editing based methods
Traditional image captioning methods predict the target sentences word by word in a one-pass manner. However, such a single-pass decoding process can neither correct the previously generated words nor take a global semantic perception. Inspired by the post-editing [24], some works take advantage of a two-pass decoding framework for caption generation. Post-editing based image captioning methods generate refined captions based on the first-pass generated captions and the image features. Guo et al. [25] proposed a Ruminant Network, in which a ruminant decoder refined the first-pass generated captions by a base decoder to re-generate new comprehensive captions. Sammani et al. [26] further employed a copy mechanism to make the best of the existing caption. The proposed method can copy or fix the details of existing captions. Most recently, Song et al. [27] presented a Context-Aware Auxiliary Guidance mechanism (CAAG) to provide global semantic guidance for the second-pass generation process. In this framework, the pre-generated caption is encoded as semantic features, and CAAG performs attention to these features to select useful information for the refined-caption generation.
2.6. Vision-language pre-training

Inspired by the success of pre-training models, Vision-Language Pre-training (VLP) has been proposed. Zhou et al. [28] presented a unified VLP model that can be fine-tuned for visual-language generation and understand tasks, which was pre-trained on a large amount of image-text pairs. It's verified that the pre-training stage can significantly speed up the learning and improve model performance of image captioning. But, it simply concatenates object features and image region features and uses self-attention to learn alignments between these features. Thus Li et al. [29] proposed to use object tags of images as anchor points to ease the alignment learning. It achieves new state-of-the-art performance on image captioning. Methods use image-caption pairs for pre-training fail to generate novel object captions, thus Hu et al. [30] broke this dependency and pre-trained visual-language alignments based on image-tag pairs, boosting image captioning on novel object captioning and the general task.

3. Datasets and evaluation metrics

3.1. Datasets

| Table 1. Publicly used datasets for image captioning. |
|---------------------------------|---------|------|--------|--------|--------|
|                                | Train   | Validate | Test | Public split | Online test | Captions |
|--------------------------------|---------|-----------|------|---------------|-------------|----------|
| Flickr8k                       | 6,000   | 1,000     | 1,000 | √              | ×           | 5        |
| Flickr30k                      | 29,783  | 1,000     | 1,000 | √              | ×           | 5        |
| MS COCO                       | 113,287 | 5,000     | 5,000 | √              | 40,775      | 5        |

Flickr8k [31], Flickr30k [32], and MS COCO [33] are widely used datasets for image captioning. We compare these datasets in Table 1.

- **Flickr8k.** A dataset that contains 8,000 images and each is annotated with five natural language descriptions based on crowd-sourcing service from Amazon Mechanical Turk. The images are mainly about humans and animals, and each sentence provides the salient entities and events of the corresponding image.

- **Flickr30k.** An extension of Flickr8k which contains 31,783 images in total. It further provides comprehensive ground-truth correspondences between image regions and caption phrases.

- **MS COCO.** The most popular dataset for image captioning, which has 123,287 images in total. Each image is annotated with at least five caption sentences. It also provides an online test service with an online test split which contains 40,775 images. Captions of the online test split are not available publicly.

3.2. Evaluation metrics

The widely used automatic evaluation metrics contains BLEU [34], METEOR [35], CIDEr [36], ROUGE [37] and SPICE [38]. Note that, higher scores indicate better sentences.

- **BLEU** measures the matching degree of n-gram phrases in the generated sentence and the corresponding ground truth sentence (annotated by humans). Compared with BLEU, METEOR makes better semantic correlation and is more relevant to human judgments. Specifically, it uses WordNet to take stems and synonyms into consideration. CIDEr performs human consensus. It uses the term frequency-inverse document frequencies (TF-IDF) of the generated and the ground truth sentences to calculate the similarity score (cosine distance). ROUGE measures the overlapping units such as word sequences, word pairs, and n-grams between the generated sentence and the ground truth sentence. SPICE is calculated based on semantic concepts. It transforms both the generated sentence and the ground truth sentence into an intermediate representation and measures how well the generation model recovers objects, attributes, and the relations between them.
4. Future research directions
In recent years, image captioning has achieved great success. However, in our viewpoint, there are still some directions worthy to be explored. First, existing methods rely heavily on caption-annotated images, lacking the use of those without annotations. Therefore, unsupervised learning will be a way to solve this problem. Second, most current approaches generate captions of images, but the caption phrases are not well aligned to the image elements. More grounded image captions should be further explored. Last but not least, external knowledge is a good point to probe, since it can help the model to generate more informative sentences and enhances knowledge reasoning in the generated caption.

5. Conclusion
In this paper, we review the advanced image captioning methods and classify these methods into different categories. Representative methods of each category are discussed. We also give a brief introduction of the evaluation metrics and widely used datasets. Furthermore, we briefly discuss the future research directions of image captioning.

6. References
[1] Tan, Huibin, et al. Nonnegative constrained graph based canonical correlation analysis for multi-view feature learning. Neural Processing Letters 50.2 (2019): 1215-1240.
[2] Dong, G., Zhang, X., Lan, L., Wang, S., & Luo, Z. (2019). Label guided correlation hashing for large-scale cross-modal retrieval. Multimedia Tools and Applications, 78(21), 30895-30922.
[3] Vinyals, O., Toshev, A., Bengio, S. and Erhan, D, Show and tell: A neural image caption generator, in CVPR 2015, pp. 3156-3164.
[4] X. Jia, E. Gavves, B. Fernando, and T. Tuyltaelars, Guiding the long-short term memory model for image caption generation, in ICCV 2016, pp. 2407–2415.
[5] C. Wang, H. Yang, C. Bartz, and C. Meinel, Image captioning with deep bidirectional lstms, in ACM MM 2016, pp. 988–997.
[6] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, Computer Science, pp. 2048–2057, 2015.
[7] J. Jin, K. Fu, R. Cui, F. Sha and C. Zhang, Aligning where to see and what to tell: image caption with region-based attention and scene factorization, arXiv:1506.06272, 2015.
[8] M. Pedersoli, T. Lucas, C. Schmid, and J. Verbeek, Areas of attention for image captioning, in ICCV 2017, pp. 1242–1250.
[9] J. Lu, C. Xiong, D. Parikh, and R. Socher, Knowing when to look: Adaptive attention via a visual sentinel for image captioning, in CVPR 2017, vol. 6, p. 2.
[10] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, Bottom-up and top-down attention for image captioning and vqa, arXiv:1707.07998, 2017.
[11] L. Huang, W. Wang, Y. Xia, and J. Chen, Adaptively aligned image captioning via adaptive attention time, in NIPS 2019, pp. 8942–8951.
[12] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, Attention on attention for image captioning, in ICCV 2019, pp. 4634–4643.
[13] Q. Wu, C. Shen, P. Wang, A. Dick, and A. van den Hengel, Image captioning and visual question answering based on attributes and external knowledge, in TPAMI, 2017.
[14] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, Boosting image captioning with attributes, in ICCV 2017, pp. 22–29.
[15] J. Zhang, K. Mei, Y. Zheng and J. Fan, Integrating part of speech guidance for image captioning, in TMM, vol. 23, pp. 92–104, 2020.
[16] Y. Huang, J. Chen, W. Ouyang, W. Wan, and Y. Xue, Image captioning with end-to-end attribute detection and subsequent attributes prediction, in TIP, vol. 29, pp. 4013–4026, 2020.
[17] S. Chen, Q. Jin, P. Wang, and Q. Wu, Say as you wish: Fine-grained control of image caption generation with abstract scene graphs, in CVPR, 2020, pp. 9962–9971.
[18] X. Yang, K. Tang, H. Zhang, and J. Cai, Auto-encoding scene graphs for image captioning, in CVPR 2019, pp. 10 685–10 694.
[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, Attention is all you need, in *NIPS* 2017, pp. 5998–6008.
[20] G. Li, L. Zhu, P. Liu and Y. Yang, Entangled transformer for image captioning, in *ICCV* 2019, pp. 8928–8937.
[21] Y. Pan, T. Yao, Y. Li, and T. Mei, X-linear attention networks for image captioning, in *CVPR* 2020, pp. 10 971–10 980.
[22] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, Meshed-memory transformer for image captioning, in *CVPR* 2020, pp. 10 578–10 587.
[23] J. Ji, Y. Luo, X. Sun, F. Chen, G. Luo, Y. Wu, Y. Gao, and R. Ji, Improving image captioning by leveraging intra-and inter-layer global representation in transformer network, in *AAAI* 2021.
[24] R. Chatterjee, M. Weller, M. Negri, and M. Turchi, Exploring the planet of the apes: a comparative study of state-of-the-art methods for mt automatic post-editing, in *ACL* (Volume 2: Short Papers), 2015, pp. 156–161.
[25] L. Guo, J. Liu, S. Lu and H. Lu, Show, tell, and polish: Ruminant decoding for image captioning, in *TMM*, vol. 22, no. 8, pp. 2149–2162, 2019.
[26] F. Sammani and L. Melas-Kyriazi, Show, edit and tell: A framework for editing image captions, in *CVPR* 2020, pp. 4808–4816.
[27] Z. Song, X. Zhou, Z. Mao, and J. Tan, Image captioning with context-aware auxiliary guidance, *arXiv*:2012.05545, 2020.
[28] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, and J. Gao, Unified vision-language pre-training for image captioning and vqa, in *AAAI* 2020, vol. 34, no. 07, pp. 13 041–13 049.
[29] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei et al., Oscar: Object-semantics aligned pre-training for vision-language tasks, in *ECCV* 2020, pp. 121–137.
[30] X. Hu, X. Yin, K. Lin, L. Wang, L. Zhang, J. Gao, and Z. Liu, Vivo: Surpassing human performance in novel object captioning with visual vocabulary pre-training, *arXiv*:2009.13682, 2020.
[31] M. Hodosh, P. Young, and J. Hockenmaier, Framing image description as a ranking task: Data, models and evaluation metrics, in *JAIR*, vol. 47, pp. 853–899, 2013.
[32] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models, in *ICCV*, 2015, pp. 2641–2649.
[33] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, Microsoft coco: Common objects in context, in *ECCV*. Springer, 2014, pp. 740–755.
[34] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, Bleu: a method for automatic evaluation of machine translation, in *ACL* 2002, pp. 311–318.
[35] S. Banerjee and A. Lavie, Meteor: An automatic metric for mt evaluation with improved correlation with human judgments, in *ACL workshop*, 2005, pp. 65–72.
[36] R. Vedantam, C. L. Zitnick and D. Parikh, Cider: Consensus-based image description evaluation, in *CVPR* 2015, pp. 4566–4575.
[37] C. Flick, Rouge: A package for automatic evaluation of summaries, in *The Workshop on Text Summarization Branches Out*, 2004, p. 10.
[38] P. Anderson, B. Fernando, M. Johnson and S. Gould, Spice: Semantic propositional image caption evaluation, in *ECCV*. Springer, 2016, pp. 382–398

**Acknowledgments**

The research is supported by the National Natural Science Foundation of China under Grant No. 62002373.