SIP (2020), vol. 9, e11, page 1 of 8 © The Authors, 2020.

doi:10.1017/ATSIP.2020.10

INDUSTRIAL TECHNOLOGY ADVANCES

Vision and language: from visual perception to content creation

TAO MEI, WEI ZHANG AND TING YAO

Vision and language are two fundamental capabilities of human intelligence. Humans routinely perform tasks through the interactions between vision and language, supporting the uniquely human capacity to talk about what they see or hallucinate a picture on a natural-language description. The valid question of how language interacts with vision motivates us researchers to expand the horizons of computer vision area. In particular, "vision to language" is probably one of the most popular topics in the past 5 years, with a significant growth in both volume of publications and extensive applications, e.g. captioning, visual question answering, visual dialog, language navigation, etc. Such tasks boost visual perception with more comprehensive understanding and diverse linguistic representations. Going beyond the progresses made in "vision to language," language can also contribute to vision understanding and offer new possibilities of visual content creation, i.e. "language to vision." The process performs as a prism through which to create visual content conditioning on the language inputs. This paper reviews the recent advances along these two dimensions: "vision to language" and "language to vision." More concretely, the former mainly focuses on the development of image/video captioning, as well as typical encoder–decoder structures and benchmarks, while the latter summarizes the technologies of visual content creation. The real-world deployment or services of vision and language are elaborated as well.

Keywords: Deep learning, Computer vision, Artificial intelligence

Received 1 November 2019; Revised 28 February 2020

I. INTRODUCTION

Computer vision (CV) and natural language processing (NLP) are two most fundamental disciplines under a broad area of artificial intelligence (AI). CV is regarded as a field of research that explores the techniques to teach computers to see and understand the digital content such as images and videos. NLP is a branch of linguistics that enables computers to process, interpret, and even generate human language. With the rise and development of deep learning over the past decade, there has been a steady momentum of innovation and breakthroughs that convincingly push the limits and improve the state-of-the-art of both vision and language modeling. An interesting observation is that the research in the two area starts to interact and many previous experiences have shown that by doing so can naturally build up the circle of human intelligence.

In general, the interactions between vision and language have proceeded along two dimensions: vision to language and language to vision. The former predominantly recognizes or describes the visual content with a set of individual words or a natural sentence in the form of tags [1], answers [2], captions [3–5], and comments [6]. For example, a tag usually denotes a specific object, action, or event in visual content. An answer is a response to a question about the details depicted in an image or a video. A caption goes beyond tags or answers by producing a natural-language utterance (usually a sentence) and a comment is also a sentence which expresses an emotional state on visual content. The latter of language to vision basically generates visual content according to natural language inputs. One typical application is to create an image or a video from text. For instance, given a textual description of "this small bird has short beak and dark stripe down the top, the wings are a mix of brown, white, and black," the goal of text-to-image synthesis is to generate a bird image which meets all the details.

This paper reviews the recent state-of-the-art advances of AI technologies which boost both vision to language, particularly image/video captioning, and language to vision. The real-world deployments in the two fields are also presented as the good examples of how AI transforms the customer experiences and enhances user engagement in industrial applications. The remaining sections are organized as follows. Section II describes the development of vision to language by outlining a brief road map of key technologies on image/video captioning, distilling a typical encoder–decoder structure, and summarizing the evaluations on a popular benchmark. The practical applications of

JD AI Research, Building A, North-Star Century Center, 8 Beichen West Road, Beijing, China

Corresponding author:
Tao Mei
Email: tmei@jd.com

vision to language are further presented. Section III details the technical advancements on language to vision in terms of different conditions and strategies for generation, followed by a summary of progresses on language to image, language to video, and AI-empowered applications. Finally, we conclude the paper in Section IV.

II. VISION TO LANGUAGE

This section summarizes the development of vision to language (particularly image/video captioning) in several aspects, ranging from the road map of key techniques and benchmarks, typical encoder–decoder architectures, to the evaluation results of representative methods.

A) Road map of vision to language

In the past 10 years, we have witnessed researchers strived to push the limits of vision to language systems (e.g. image/video captioning). Figure 1 depicts the road map for the techniques behind vision (image/video) to language and the corresponding benchmarks. Specifically, the year of 2015 is actually a watershed in captioning. Before that, the main stream of captioning is a template-based method [14,15] in image domain. The basic idea is to detect the objects or actions in an image and integrate these words into predefined sentence templates as subjective, verb, and objective. At that time, most of the image captioning datasets are ready to use, such as Flickr30K and MSCOCO. At the year 2015, deep learning-based image captioning models are first presented. The common design [13] is to employ a Convolutional Neural Network (CNN) as an image encoder to produce image representations and exploit a decoder of Long Short-Term Memory (LSTM) to generate the sentence. The attention mechanism [16] is also proposed at that year which locates the most relevant spatial regions when predicting each word. After that, the area of image captioning is growing very fast. Researchers came up with a series of innovations, such as augmenting image features with semantic attributes [3] or visual relations [4], predicting novel objects through leveraging unpaired training data [17,18], and even going a step further to perform language navigation [19]. Another extension direction of captioning in image domain is to produce multiple sentences or phrases for an image, aiming to recapitulate more details within image. In between, dense image captioning [20] and image paragraph generation [21] are typical ones, which generate a set of descriptions or paragraph that describes image in a finer fashion.

The start point of captioning in video domain is also in the year of 2015. Then, researchers start to remould the CNN plus RNN captioning framework toward the scenario of captioning in video domain. A series of techniques (e.g. temporal attention, embedding, or attributes) are explored to further improve video captioning. Concretely, Yao *et al.*'s technique [22] is one of the early attempts that incorporates temporal attention mechanism into captioning framework by learning to attend to the most relevant frames at each decoding time step. Pan *et al.* [23] integrate LSTM with semantic embedding to preserve the semantic relevance between video content and the entire sentence. Pan *et al.* [24] further augment captioning model to emphasize the detected visual attributes in the generated sentence. It is also worthy mentioned that in 2016, MSR-VTT video captioning dataset [25] is released which has been widely used and already downloaded by more than 100 groups worldwide. Most recently, Aafaq *et al.* [26] apply short Fourier transform across all the frame-level features along the temporal dimension to fuse all frame-level features into video-level representation and further enhance video captioning. Another recent attempt for video captioning is to speed up the training procedure by fully employing convolutions

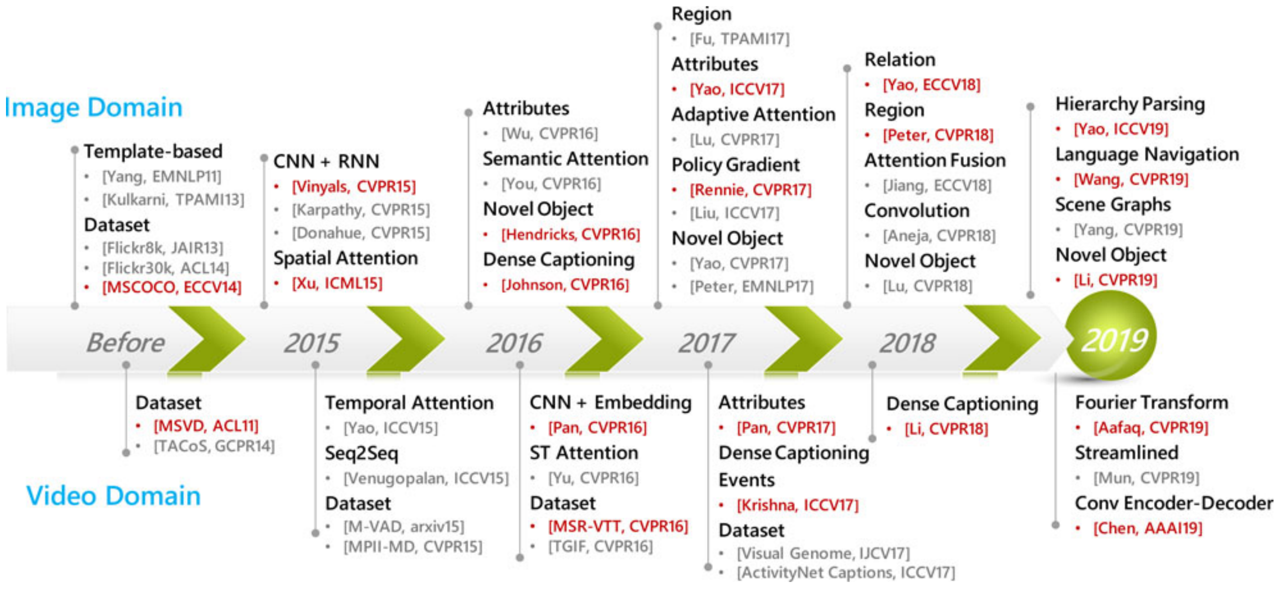
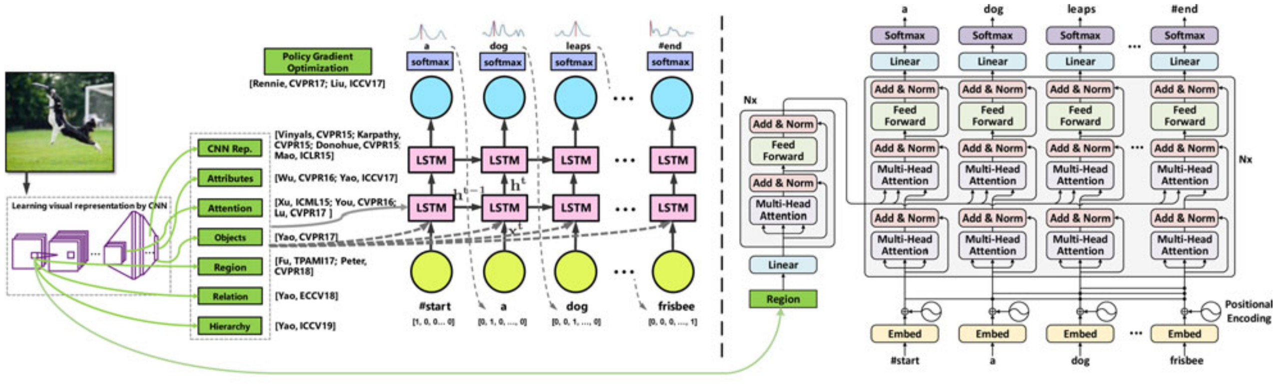


Fig. 1. A road map for the techniques and datasets in vision (image/video) to language in 10 years.



(a) CNN encoder plus LSTM decoder

(b) Transformer-based encoder-decoder

Fig. 2. The typical architectures of (a) CNN encoder plus LSTM decoder and (b) transformer-based encoder–decoder for image captioning.

in both encoder and decoder networks [27]. Nevertheless, considering that videos in real life are usually long and contain multiple events, the conventional video captioning methods generating only one caption for a video in general will fail to recapitulate all the events in the video. Hence the task of dense video captioning [28,29] is introduced recently and the ultimate goal is to generate a sentence for each event occurring in the video.

B) Typical architectures

According to the road map of vision to language, the mainstream of modern image captioning follows the structure of CNN encoder plus LSTM decoder, as shown in Fig. 2(a). In particular, given an image, image features can be firstly extracted through multiple ways: (1) directly taking the outputs of fully-connected layers as image features [13]; (2) incorporating high-level semantic attributes into image features [3]; (3) performing attention mechanism to measure the contribution of each image region [16]; (4) extracting region-level features [2] and further exploring relation [4] or image hierarchy [5] on the region-level features. The image features will be further fed into LSTM decoder to generate the output sentence, one word at each time step. In the training stage, the next word is generated based on the previous ground-truth words while during testing the model uses the previously generated words to predict the next word. In order to bridge the mismatch between training and testing, reinforcement learning [8,9,30] is usually exploited to directly optimize LSTM decoder with the sentence-level reward, such as CIDEr or METEOR.

Taking the inspiration from the recent successes of Transformer self-attention networks [31] in machine translation, recent attention has been geared toward exploring Transformer-based structure [32] in image captioning. Figure 2(b) depicts the typical architecture of Transformer-based encoder–decoder. Different from CNN encoder plus LSTM decoder that capitalizes on LSTM to model word dependency, Transformer-based encoder–decoder model fully utilizes attention mechanism to capture the global dependencies among inputs. For encoder, N multi-head self-attention layers are stacked to model the self-attention among input image regions. The decoder contains a stack of N multi-head attention layers, each of which consists of a self-attention sub-layer and a cross-attention sub-layer. More specifically, the self-attention sub-layer is firstly adopted to capture word dependency and the cross-attention sub-layer is further utilized to exploit the co-attention across vision (image regions from encoder) and language (input words).

Similar to the mainstream in image captioning, the typical paradigm in video captioning is also essentially an encoder–decoder structure. A video is first encoded into a set of frame/clip/shot features via 2D CNN [13] or 3D CNN [33,34]. Next, all the frame-level, clip-level or shot-level visual features are fused into video-level representations through pooling [23], attention [22], or LSTM-based encoder [35]. The video-level features are then fed into LSTM decoder to produce a natural sentence.

C) Evaluation and applications

Evaluation. Here we summarize the reported performance of representative image captioning methods on the testing server of popular benchmark COCO [36] in Table 1. In terms of all the evaluation metrics, GCN-LSTM [4] and HIP [5] lead to performance boost against other captioning systems, which verifies the advantage of exploring relations and hierarchal structure among image regions.

Applications. Recently, there exist several emerging applications which involve the technology of vision to language. For example, captioning is integrated into online chatbot [37,38] and an ai-created poetry [39] is published in China. In JD.com, we utilize captioning techniques for personalized product description generation last year, which aims to produce compelling recommendation reasons for billions of products automatically.

III. LANGUAGE TO VISION

This section discusses from another direction of "language to vision," i.e. visual content generation guided by language inputs. In this section, we start by reviewing the road map development, as well as the technical

Table 1. The reported performance (%) of image captioning on COCO testing server with 5 reference captions (c5) and 40 reference captions (c40).

		B@4		METEOR		ROUGE-L		CIDEr-D	
Model	Group	c5	c40	c5	c40	c5	c40	c5	c40
HIP [5]	JD AI, ICCV'19	**39.3**	**71**	**28.8**	**38.1**	**59**	**74.1**	**127.9**	**130.2**
GCN-LSTM [4]	JD AI, ECCV'18	38.7	69.7	28.5	37.6	58.5	73.4	125.3	126.5
RFNet [7]	Tencent, ECCV'18	38	69.2	28.2	37.2	58.2	73.1	122.9	125.1
Up-Down [2]	MSR, CVPR'18	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
LSTM-A [3]	MSRA, ICCV'17	35.6	65.2	27	35.4	56.4	70.5	116	118
Watson Multimodal [8]	IBM, CVPR'17	35.2	64.5	27.0	35.5	56.3	70.7	114.7	116.7
G-RMI [9]	Google, ICCV'17	33.1	62.4	25.5	33.9	55.1	69.4	104.2	107.1
MetaMind/VT-GT [10]	Salesforce, CVPR'17	33.6	63.7	26.4	35.9	55	70.5	104.2	105.9
reviewnet [11]	CMU, NIPS'16	31.3	59.7	25.6	34.7	53.3	68.6	96.5	96.9
ATT [12]	Rochester, CVPR'16	31.6	59.9	25	33.5	53.5	68.2	94.3	95.8
Google [13]	Google, CVPR'15	30.9	58.7	25.4	34.6	53	68.2	94.3	94.6

advancements in this area. Then we discuss the open issues and applications particularly from the perspective of industry.

Visual Content Generation. We briefly introduce the domain of visual generation, since "language to vision" is deeply rooted in the same techniques. Over the past few years, we have witnessed great progresses in visual content generation. The origin of visual generation dates back to [40], where multiple networks are jointly trained in an adversarial manner. Subsequent works generate images in specific domains such as face [41–43], person [44–46], as well as generic domains [47,48]. From the perspective of inputs, the generation can also be treated as conditioning on different information, e.g. noise vector [40], semantic label [49], textual captions [50], scene-graph [51], and images [52,53]. Among all these works, visual generation based on natural languages plays one of the most promising branches, since semantics are directly incorporated into the pixel-wise generation process.

A) Road map of language to vision

Figure 3 summarizes recent development of "language to vision." In general, both the vision and language modalities are becoming more and more complicated, and the results are much more visually convincing, compared to when it was firstly introduced in 2014.

The fundamental architecture is based on a conditional generative adversarial network, where the conditioning input is usually the encoded natural language. After a series of transposed-convolutions, the language input is gradually mapped to a visual image with higher and higher resolution. The key challenges are in two folds: (1) how to interpret the language input, i.e. language representation, and (2) how to align the visual and textual modalities, i.e. the semantic consistency between vision and language. Recent results on single object (bottom) have already been visually plausible to human perception. However, state-of-the-art models are still struggling in generating scenes with multiple objects interacting with each other.

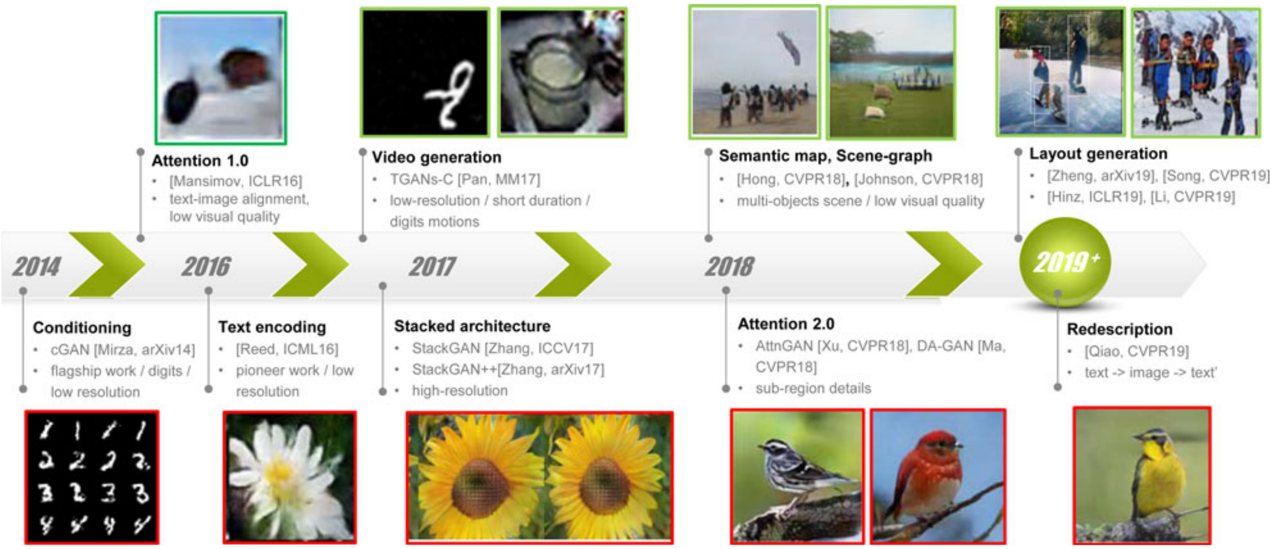

Attention 1.0
• [Mansimov, ICLR16]
• text-image alignment, low visual quality
Video generation
• TGANs-C [Pan, MM17]
• low-resolution / short duration / digits motions
Semantic map, Scene-graph
• [Hong, CVPR18], [Johnson, CVPR18]
• multi-objects scene / low visual quality
Layout generation
• [Zheng, arXiv19], [Song, CVPR19]
• [Hinz, ICLR19], [Li, CVPR19]
2014
2016
2017
2018
2019+
Conditioning
• cGAN [Mirza, arXiv14]
• flagship work / digits / low resolution
Text encoding
• [Reed, ICML16]
• pioneer work / low resolution
Stacked architecture
• StackGAN [Zhang, ICCV17]
• StackGAN++[Zhang, arXiv17]
• high-resolution
Attention 2.0
• AttnGAN [Xu, CVPR18], DA-GAN [Ma, CVPR18]
• sub-region details
Redescription
• [Qiao, CVPR19]
• text -> image -> text'

Fig. 3. The road map of "language-to-vision" in past five years, while milestone techniques are marked along the year axis. Top: single object generation. Bottom: multiple-objects scene generation.

B) Technical advancements

The success in language to vision generation is mostly based on the following technical advancements, which have become standard practices commonly accepted by the research community.

Conditioning Input. Following the standard GAN framework [40,49] derived the conditional version GAN, which allows visual generation according to language inputs. The conditioning information can be in any form of language, such as tag, sentence, paragraph, image, scene-graph, and layout. Almost all subsequent works in "language to vision" are based on the conditioning architecture. However at that time, only MNIST [54] digits are demonstrated in low resolution, and the conditioning information is merely a digit-label.

Text Encoding. GAN-INT-CLS [50] is the first work based on natural-language inputs. For the first time, it bridges the gap from natural language sentences to image pixels. The key step is based on learning a text representation based on a recurrent network to capture visual clues. The rest is mostly following [49]. Additionally, a matching-aware discriminator is proposed to keep the consistency between the generated image and textual input. Though the results still look primitive, people can draw flower images by altering textual inputs.

Stacked Architecture. Another big advancement is by stackGAN [55,56], where stacked generators are introduced for high-resolution image generation. Different from previous works, stackGAN can generate realistic 256 × 256-pixel images by decomposing the generator into multiple stages stacked sequentially. The Stage-I network only sketches the primitive shape and color of the object based on text representation, yielding a low resolution image. The Stage-II network further fills details, such as textures, conditioning on the Stage-I result. A conditioning augmentation technique is also introduced to augment the textual input and stabilize the training process. Compared to [50], the visual quality is much improved based on this stacked architecture. Similar idea is also adopted in Progressively-Growing GAN [42].

Attention Mechanism. As in other vision tasks, attention is effective in highlighting key information. In "language to vision," attention is particularly useful in aligning keywords (language) and image patches (vision) during the generation process. Two generations (v1.0 and v2.0) of attention basically follow this paradigm, but differs in many details, e.g. network architecture, text encoding. Attention 1.0, AlignDraw [57], proposes to iteratively paint on a canvas by looking at different words at different stages. However, the results were not promising at that time. Attention 2.0, AttnGAN [58] and DA-GAN [59], basically follows the similar paradigm, but improves significantly on image quality, e.g. fine-grained details.

Semantic Layout. Recent studies [46,60,61] have demonstrated the importance of semantic layout in image generation, where layout acts as the blue-print to guide the generation process. In language to vision, semantic layout and scene-graph are introduced to reshape the language input with more semantics. Hong *et al.* [62] propose to generate object bounding-boxes first, and then refine by estimating appearances inside each box. Johnson *et al.* [51] encode objects relationship from scene graph to construct the layout for decoder generation with graph convolutions. Zheng *et al.* [63] introduce spatial constraint module and contextual fusion module to model the relative scale and offset among objects for commonsense layout generation, and Hinz *et al.* [64] further propose an object pathway for multi-objects generation with complex spatial layouts.

C) Progress and applications

The development of "Language to Vision" can be summarized as follows. On one hand, the language description is becoming more complex, i.e. from simple words to long sentences. On the other hand, the vision part is also becoming more complex, where objects-interaction and fine-grained detail are expected:

- Language: label → sentence → paragraph → scene graph
- Vision: single object → multiple objects

Language to Image. Early studies mainly focus on simple words and single-object images, e.g. birds [65], flowers [66], and generic objects [67]. As shown in Fig. 3 (bottom), the visual quality is much improved over the past few years, and some results are plausible enough to deceive human eyes.

Though single-object image can be well generated, multi-objects scene still struggles for realistic results, as in Fig. 3 (top). A general trend is to reduce the complexity by introducing semantic layout as an intermediate representation. Roughly, machines now can generate spatially reasonable images, but fine-grained details are still far from satisfactory at current stage.

Language to Video. Compared to image, language to video is more challenging due to huge volume of information and extra temporal constraint. There is only a few works studying this area. For example, Pan *et al.* [68] attempt to generate video out of captions based on 3D convolution operation. However, the results are quite limited for practical applications.

Applications. The application of "language to vision" can be roughly grouped into two categories: generation for human eyes or for machines. In certain domains (e.g. face), language to vision already starts to produce highly plausible results with industrial standards[1]. For example, people can generate royalty-free facial photos on demand[2] for games [69] or commercials, by manually specifying gender, hair, eyes. Another direction is generating data for machine and algorithms. For example, NVIDIA [70] proposed a large-scale synthetic dataset (DG-Market) for training person Re-ID models. Also some image recognition and segmentation models start to benefit from machine-generated training

[1] https://thispersondoesnotexist.com/

[2] https://github.com/SummitKwan/transparent_latent_gan

images. However, it is worth noting that despite the promising results, there is still a large gap for massive deployment in industrial products.

IV. CONCLUSION

Vision and language are two fundamental systems of human representation. Integrating the two in one intelligent system has long been an ambition in AI field. As we have discussed in the paper, on one hand, vision to language is capable of understanding visual content and automatically producing a natural-language description, and on the other hand, language to vision is able to characterize the intrinsic structure in vision data and create visual content according to the language inputs. Such interactions, while still at the early stage, motivate us to understand the mechanisms in connecting vision and language, reshape real-world applications, and re-think the end result of the integration.

REFERENCES

[1] Yao T.; Mei T.; Ngo C.-W.; Li S.: Annotation for free: video tagging by mining user search behavior, in *Proceedings of the 21st ACM International Conference on Multimedia*, Barcelona, Spain, 2013, 977–986.

[2] Anderson P. *et al.*: Bottom-up and top-down attention for image captioning and visual question answering, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, USA, 2018, 6077–6086.

[3] Yao T.; Pan Y.; Li Y.; Qiu Z.; Mei T.: Boosting image captioning with attributes, in *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, 2017, 4894–4902.

[4] Yao T.; Pan Y.; Li Y.; Mei T.: Exploring visual relationship for image captioning, in *Proceedings of the European Conference on Computer Vision*, Munich, Germany, 2018, 684–699.

[5] Yao T.; Pan Y.; Li Y.; Mei T.: Hierarchy parsing for image captioning, in *Proceedings of the IEEE International Conference on Computer Vision*, Seoul, Korea, 2019, 2621–2629.

[6] Li Y.; Yao T.; Mei T.; Chao H.; Rui Y.: Share-and-chat: achieving human-level video commenting by search and multi-view embedding, in *Proceedings of the 24th ACM International Conference on Multimedia*, Amsterdam, The Netherlands, 2016, 928–937.

[7] Jiang W.; Ma L.; Jiang Y.-G.; Liu W.; Zhang T.: Recurrent fusion network for image captioning, in *Proceedings of the European Conference on Computer Vision*, Munich, Germany, 2018, 499–515.

[8] Rennie S.J.; Marcheret E.; Mroueh Y.; Ross J.; Goel V.: Self-critical sequence training for image captioning, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, 2017, 7008–7024.

[9] Liu S.; Zhu Z.; Ye N.; Guadarrama S.; Murphy K.: Improved image captioning via policy gradient optimization of spider, in *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, 2017, 873–881.

[10] Lu J.; Xiong C.; Parikh D.; Socher R.: 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*, Honolulu, USA, 2017, 375–383.

[11] Yang Z.; Yuan Y.; Wu Y.; Cohen W.W.; Salakhutdinov R.R.: Review networks for caption generation, in *Advances in Neural Information Processing Systems*, Barcelona, Spain, 2016, 2361–2369.

[12] You Q.; Jin H.; Wang Z.; Fang C.; Luo J.: Image captioning with semantic attention, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, 2016, 4651–4659.

[13] Vinyals O.; Toshev A.; Bengio S.; Erhan D.: Show and tell: a neural image caption generator, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, USA, 2015, 3156–3164.

[14] Kulkarni G. *et al.*: Babytalk: understanding and generating simple image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.*, **35** (12) (2013), 2891–2903.

[15] Yang Y.; Teo C.L.; Daumé III H.; Aloimonos Y.: Corpus-guided sentence generation of natural images, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, UK, 2011, 444–454.

[16] Xu K. *et al.*: Show, attend and tell: neural image caption generation with visual attention, in *International Conference on Machine Learning*, Lille, France, 2015, 2048–2057.

[17] Li Y.; Yao T.; Pan Y.; Chao H.; Mei T.: Pointing novel objects in image captioning, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, USA, 2019, 12497–12506.

[18] Yao T.; Pan Y.; Li Y.; Mei T.: Incorporating copying mechanism in image captioning for learning novel objects, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, 2017, 6580–6588.

[19] Wang X. *et al.*: Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, USA, 2019, 6629–6638.

[20] Johnson J.; Karpathy A.; Fei-Fei L.: Densecap: fully convolutional localization networks for dense captioning, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, 2016, 4565–4574.

[21] Wang J.; Pan Y.; Yao T.; Tang J.; Mei T.: Convolutional auto-encoding of sentence topics for image paragraph generation, in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, Macao, China, 2019, 940–946.

[22] Yao L. *et al.*: Describing videos by exploiting temporal structure, in *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, Chile, 2015, 4507–4515.

[23] Pan Y.; Mei T.; Yao T.; Li H.; Rui Y.: Jointly modeling embedding and translation to bridge video and language, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, 2016, 4594–4602.

[24] Pan Y.; Yao T.; Li H.; Mei T.: Video captioning with transferred semantic attributes, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, 2017, 6504–6512.

[25] Xu J.; Mei T.; Yao T.; Rui Y.: Msr-vtt: a large video description dataset for bridging video and language, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, 2016, 5288–5296.

[26] Aafaq N.; Akhtar N.; Liu W.; Gilani S.Z.; Mian A.: Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, USA, 2019, 12487–12496.

[27] Chen J.; Pan Y.; Li Y.; Yao T.; Chao H.; Mei T.: Temporal deformable convolutional encoder-decoder networks for video captioning, in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. **33**, Honolulu, USA, 2019, 8167–8174.

[28] Krishna R.; Hata K.; Ren F.; Fei-Fei L.; Niebles J.C.: Dense-captioning events in videos, in *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, 2017, 706–715.

[29] Li Y.; Yao T.; Pan Y.; Chao H.; Mei T.: Jointly localizing and describing events for dense video captioning, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, USA, 2018, 7492–7500.

[30] Ren Z.; Wang X.; Zhang N.; Lv X.; Li L.-J.: Deep reinforcement learning-based image captioning with embedding reward, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, 2017, 290–298.

[31] Vaswani A. *et al.*: Attention is all you need, in *Advances in Neural Information Processing Systems*, Long Beach, USA, 2017, 5998–6008.

[32] Sharma P.; Ding N.; Goodman S.; Soricut R.: Conceptual captions: a cleaned, hypernymed, image alt-text dataset for automatic image captioning, in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, Australia, 2018, 2556–2565.

[33] Qiu Z.; Yao T.; Mei T.: Learning spatio-temporal representation with pseudo-3d residual networks, in *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, 2017, 5533–5541.

[34] Tran D.; Bourdev L.; Fergus R.; Torresani L.; Paluri M.: Learning spatiotemporal features with 3d convolutional networks, in *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, Chile, 2015, 4489–4497.

[35] Venugopalan S.; Rohrbach M.; Donahue J.; Mooney R.; Darrell T.; Saenko K.: Sequence to sequence-video to text, in *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, Chile, 2015, 4534–4542.

[36] Lin T.-Y. *et al.*: Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[37] Pan Y.; Qiu Z.; Yao T.; Li H.; Mei T.: Seeing bot, in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Tokyo, Japan, 2017, 1341–1344.

[38] Tran K. *et al.*: Rich image captioning in the wild, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition workshops*, Las Vegas, USA, 2016, 49–56.

[39] Zhou L.; Gao J.; Li D.; Shum H.-Y.: The design and implementation of xiaoice, an empathetic social chatbot, 2018. *arXiv preprint arXiv:1812.08989*.

[40] Goodfellow I. *et al.*: Generative adversarial nets, in *Advances in Neural Information Processing Systems*, Montréal, Canada, 2014, 2672–2680.

[41] Chen A.; Chen Z.; Zhang G.; Mitchell K.; Yu J.: Photo-realistic facial details synthesis from single image, in *Proceedings of the IEEE International Conference on Computer Vision*, Seoul, Korea, 2019, 9429–9439.

[42] Karras T.; Aila T.; Laine S.; Lehtinen J.: Progressive growing of gans for improved quality, stability, and variation, in *International Conference on Learning Representations*, Vancouver, Canada, 2018.

[43] Karras T.; Laine S.; Aila T.: A style-based generator architecture for generative adversarial networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, USA, 2019, 4401–4410.

[44] Ma L.; Jia X.; Sun Q.; Schiele B.; Tuytelaars T.; Van Gool L.: Pose guided person image generation, in *Advances in Neural Information Processing Systems*, Long Beach, USA, 2017, 406–416.

[45] Ma L.; Sun Q.; Georgoulis S.; Van Gool L.; Schiele B.; Fritz M.: Disentangled person image generation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, USA, 2018, 99–108.

[46] Song S.; Zhang W.; Liu J.; Mei T.: Unsupervised person image generation with semantic parsing transformation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, USA, 2019, 2357–2366.

[47] Brock A.; Donahue J.; Simonyan K.: Large scale GAN training for high fidelity natural image synthesis, in *International Conference on Learning Representations*, New Orleans, USA, 2019.

[48] Lučić M.; Tschannen M.; Ritter M.; Zhai X.; Bachem O.; Gelly S.: High-fidelity image generation with fewer labels, in *International Conference on Machine Learning*, Long Beach, USA, 2019, 4183–4192.

[49] Mirza M.; Osindero S.: Conditional generative adversarial nets, 2014. *arXiv:1411.1784*.

[50] Reed S.E.; Akata Z.; Yan X.; Logeswaran L.; Schiele B.; Lee H.: Generative adversarial text to image synthesis, in *International Conference on Machine Learning*, New York City, USA, 2016, 1060–1069.

[51] Johnson J.; Gupta A.; Fei-Fei L.: Image generation from scene graphs, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, USA, 2018, 1219–1228.

[52] Isola P.; Zhu J.-Y.; Zhou T.; Efros A.A.: Image-to-image translation with conditional adversarial networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, 2017, 1125–1134.

[53] Zhu J.-Y.; Park T.; Isola P.; Efros A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks, in *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, 2017, 2223–2232.

[54] LeCun Y.; Bottou L.; Bengio Y.; Haffner P.: Gradient-based learning applied to document recognition. *Proc. IEEE*, **86** (11) (1998), 2278–2324.

[55] Zhang H. *et al.*: Stackgan++: realistic image synthesis with stacked generative adversarial networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, **41** (8) (2018), 1947–1962.

[56] Zhang H. *et al.*: Stackgan: text to photo-realistic image synthesis with stacked generative adversarial networks, in *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, 2017, 5907–5915.

[57] Mansimov E.; Parisotto E.; Ba J.L.; Salakhutdinov R.: Generating images from captions with attention, in *International Conference on Learning Representations*, San Juan, Puerto Rico, 2016.

[58] Xu T. *et al.*: Attngan: fine-grained text to image generation with attentional generative adversarial networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, USA, 2018, 1316–1324.

[59] Ma S.; Fu J.; Chen C.W.; Mei T.: Da-gan: instance-level image translation by deep attention generative adversarial networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, USA, 2018, 5657–5666.

[60] Bau D. *et al.*: Gan dissection: visualizing and understanding generative adversarial networks, in *International Conference on Learning Representations*, New Orleans, USA, 2019.

[61] Dong H.; Liang X.; Gong K.; Lai H.; Zhu J.; Yin J.: Soft-gated warping-gan for pose-guided person image synthesis, in *Advances in Neural Information Processing Systems*, Montréal Canada, 2018, 474–484.

[62] Hong S.; Yang D.; Choi J.; Lee H.: Inferring semantic layout for hierarchical text-to-image synthesis, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, USA, 2018, 7986–7994.

[63] Zheng H.; Bai Y.; Zhang W.; Mei T.: Relationship-aware spatial perception fusion for realistic scene layout generation, 2019. *arXiv:1909.00640*.

[64] Hinz T.; Heinrich S.; Wermter S.: Generating multiple objects at spatially distinct locations, in *International Conference on Learning Representations*, New Orleans, USA, 2019.

[65] Welinder P. *et al.*: Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

[66] Nilsback M.-E.; Zisserman A.: Automated flower classification over a large number of classes, in *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, Bhubaneswar, India, 2008, 722–729.

[67] Russakovsky O. *et al.*: ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis. (IJCV)*, **115** (3) (2015), 211–252.

[68] Pan Y.; Qiu Z.; Yao T.; Li H.; Mei T.: To create what you tell: generating videos from captions, in *Proceedings of the 25th ACM International Conference on Multimedia*, Mountain View, USA, 2017, 1789–1798.

[69] Shi T.; Yuan Y.; Fan C.; Zou Z.; Shi Z.; Liu Y.: Face-to-parameter translation for game character auto-creation, in *Proceedings of the IEEE International Conference on Computer Vision*, Seoul, Korea, 2019, 161–170.

[70] Zheng Z.; Yang X.; Yu Z.; Zheng L.; Yang Y.; Kautz J.: Joint discriminative and generative learning for person re-identification, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, USA, 2019, 2138–2147.

Tao Mei is a Technical Vice President with JD.com and the Deputy Managing Director of JD AI Research, where he also serves as the Director of Computer Vision and Multimedia Lab. Prior to joining JD.com in 2018, he was a Senior Research Manager with Microsoft Research Asia in Beijing, China. He has authored or co-authored over 200 publications (with 12 best paper awards) in journals and conferences. He holds over 50 US and international patents. He is or has been an Editorial Board Member of IEEE Trans. on Image Processing, IEEE Trans. on Circuits and Systems for Video Technology, IEEE Trans. on Multimedia, ACM Trans. on Multimedia, Pattern Recognition, etc. He is a Fellow of IEEE (2019), a Fellow of IAPR (2016), a Distinguished Scientist of ACM (2016), a Distinguished Industry Leader of APSIPA (2019), and a Distinguished Industry Speaker of IEEE Signal Processing Society (2017).

Wei Zhang is now a Senior Researcher in JD AI Research, Beijing, China. He received his Ph.D degree from the Department of Computer Science in the City University of Hong Kong, Hong Kong, China, in 2015. He was a visiting scholar in the DVMM group of Columbia University, New York, NY, USA, in 2014. He was in the Chinese Academy of Sciences. His research interests include computer vision, visual object analysis. He has won the runner-up in TRECVID Instance Search in 2012, the Best Demo Award in ACM-HK openday 2013. He serves as the guest editor for TOMM, co-chair for ICME workshop, MMM special session.

Ting Yao is currently a Principal Researcher in Vision and Multimedia Lab at JD AI Research, Beijing, China. His team is focusing on the research and innovation of video understanding, vision and language, and deep learning. Prior to joining JD.com, he was a Researcher with Microsoft Research Asia in Beijing, China. Dr. Yao is an active participant of several benchmark evaluations. He is the principal designer of the top-performing multimedia analytic systems in international competitions such as COCO Image Captioning, Visual Domain Adaptation Challenge 2019 & 2018 & 2017, and ActivityNet Large Scale Activity Recognition Challenge 2019 & 2018 & 2017 & 2016. His works have led to many awards, including ACM SIGMM Outstanding Ph.D. Thesis Award 2015, ACM SIGMM Rising Star Award 2019, and IEEE TCMC Rising Star Award 2019. He is also an Associate Editor of IEEE Trans. on Multimedia.