Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Perspective:

Visual knowledge: an attempt to explore machine creativity

Yueting ZHUANG‡, Siliang TANG

College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China

E-mail: yzhuang@zju.edu.cn; siliang@zju.edu.cn

Received Mar. 3, 2021; Revision accepted Apr. 21, 2021; Crosschecked Apr. 29, 2021

https://doi.org/10.1631/FITEE.2100116

1 Introduction—starting at noetic science

One question that has long puzzled the artificial intelligence (AI) community is: Can AI be creative? Or, can the reasoning process be creative? Starting at noetic science, this paper discusses the issues of visual knowledge representation and its potential applications to machine creativity. In this paper, we enumerate related research on imagery-thinking-based reasoning, then focus on a special type of visual knowledge representation, i.e., visual scene graph, and finally review the problem of visual scene graph construction and its potential applications in detail. All the evidence suggests that visual knowledge and visual thinking not only can improve the performance of current AI tasks but can be used in the practice of machine creativity.

AI has ushered in a new era of development. Existing algorithms have achieved fairly good results in clustering, classification, logical reasoning, and proving. However, looking back at the early definition of AI (McCarthy et al., 2006) which aims at machines recognizing, thinking, and learning like humans, there is still a huge gap in current algorithms, especially in terms of human-like creativity.

As early as the 1980s, an AI boom was ongoing, while its guiding concepts were still under discussion. One of China's most prominent scientists Xuesen QIAN proposed that China should establish the field of noetic science to study the laws and forms of human-like thinking activity. Noetic science is the subject of studying the relationship between consciousness and the brain, mind and matter, subjective and objective. QIAN advocated that the development of noetic science should be combined with AI and computers and that it should be addressed by the construction of abstract thinking and imagery (intuitive) thinking, social thinking, and peculiar thinking (inspirational thinking). QIAN's proposition coincides with early research in brain science in the 1960s (Gazzaniga, 1967); that is, the left brain is responsible for logical thinking, such as language, logical analysis, reasoning, abstraction, computational language memory, and writing, while the right brain is responsible for imagery thinking including intuition, emotion, graphic perception, imagery memory, art, music, vision, physical coordination, and inspiration. His ideas and suggestions broke through the mainstream framework of AI at that time, inspiring the realization of machine creativity. Even today, these ideas and suggestions still bear significant importance and theoretical value guiding future work.

In recent years, as AI has developed, mainstream research communities including top journals (such as *Science* and *Nature*) and top AI conferences (such as AAAI and IJCAI) have also begun to focus on intelligence that is capable of creativity. The core question is the simulation of creative thinking; that is, can AI be creative? Can the reasoning process be creative?

‡ Corresponding author
ORCID: Yueting ZHUANG, https://orcid.org/0000-0001-9017-2508

Take the creative behavior of advertising design as an example. It involves a large amount of visual information such as object shape, spatial relationship, color, and texture. Human designers need to reason under the guidance of imagery thinking with incomplete information. This kind of reasoning is a jumping, discontinuous thinking process, in which we humans will use "mental imagery" (Denis, 1991), the ability to arrange, combine, reconstruct, and manipulate related visual information in the brain, to explore, imagine, and reason about feasible design solutions. This process is also known as "visual thinking" (Arnheim, 1997). For machines to achieve the ability to reason and create, it would be crucial to properly preserve visual knowledge, as it serves as the basis for the algorithm to understand the visual world. The way of representing common sense and the relationships between objects in the real world is the first step for the machine to create. Current AI algorithms have made some progress in creative thinking, but mental imagery reasoning and visual thinking remain to be explored.

2 Related work on imagery-thinking-based reasoning

Related work supporting imagery-thinking-based reasoning can be traced back to case-based reasoning (CBR) in the 1980s (Kolodner, 2014). CBR is a typical paradigm of AI and cognitive science that is based on analogy. The basic idea of CBR is to simulate the process of reasoning based on the database (of cases). Its basic steps include:

1. Retrieval: Given the target problem, retrieve related cases from the database.

2. Reuse: The problem-solving scheme for the previous case is mapped to the target problem.

3. Revision: Test the new solution in the real world (or simulation) and modify it if necessary.

4. Retain: After the solution is successfully applied to the target problem, store this new experience as a new case in the database.

CBR is often used in reasoning systems such as mechanical repairs, doctors' diagnosis and treatment, and judges' decision-making. We again take advertising design as an example. Suppose there are advertising examples C_1, C_2, ..., C_m. We use $g(C_i, P_i)$ to indicate that the characteristic P_i is obtained from example C_i. The visual characteristics of the advertisement may include advertisement rendering, advertisement slogan, coloring style, and layout. Therefore, the final result of C_{new}, which is the current new design, can be described as

$$C_{\text{new}} = g(C_1, P_1) \tilde{\bigcap} g(C_2, P_2) \tilde{\bigcap} \ldots \tilde{\bigcap} g(C_m, P_m), \quad (1)$$

where $\tilde{\bigcap}$ represents a generalized operation of combination, and C_i's contribution to C_{new} is in proportion to $g(C_i, P_i)/C_{\text{new}}$. It is easy to see that the larger the m, the less similar C_i and C_{new}.

In CBR, advertising design can be abstracted as a reasoning system composed of "vision" (visual features) and "symbols" (location, combination). Imagery thinking and logical thinking are also considered to a certain extent. Although this idea originated in the 1980s, we can still see the influence of CBR on some recent papers, e.g., the best paper of *ACM Trans Multim Comput Commun Appl* published in 2017 (Yang XY et al., 2016).

In recent years, generative adversarial networks (GANs) (Radford et al., 2015) have made great progress in the field of image generation. GAN uses the zero-sum game between the discriminator and the generator to make the generation distribution fit the ground-truth data distribution. The generator obtained by a GAN can produce images that look close enough to the real images. Among the GAN models, creative adversarial networks (CANs) (Elgammal et al., 2017) produce creative paintings that pass the Turing test by adding the technique of style judgment (Fig. 1).

Fig. 1 Paintings generated by a creative adversarial network (CAN) (Elgammal et al., 2017)

Although GAN and their variants have brought significant progress to machine creativity, there are still many problems with such methods. For example, GAN is prone to the problem of mode collapse and mode drop (Bau et al., 2019). The reason for such shortcomings is that GAN is essentially distribution fitting. The lack of logical thinking and imagery thinking makes it impossible to carry out human-like innovation.

3 Visual knowledge representation and scene graph

The problems in the generation process of GAN made us realize that real human-like machine creativity requires the effective coordination of logical thinking and imagery thinking. How to coordinate two completely different ideologies under one unified framework is an urgent task for machine creativity. In 2019, Prof. Yunhe PAN proposed the theory of visual knowledge representation (Pan, 2019, 2020a) and multiple knowledge representation (Pan, 2020b). In these works, he systematically explained "visual knowledge," a new way of knowledge representation that can effectively integrate logical thinking and imagery thinking. It is believed that visual knowledge has the following characteristics:

1. It can express the spatial shape, size, and relationship of objects, as well as color and texture.

2. It can express the relationship of objects' movement, speed, and time.

3. It can perform the spatio-temporal transformation, manipulation, and reasoning of objects, including shape transformation, action transformation, speed transformation, scene transformation, various time-space analogies, association, and prediction based on spatio-temporal reasoning.

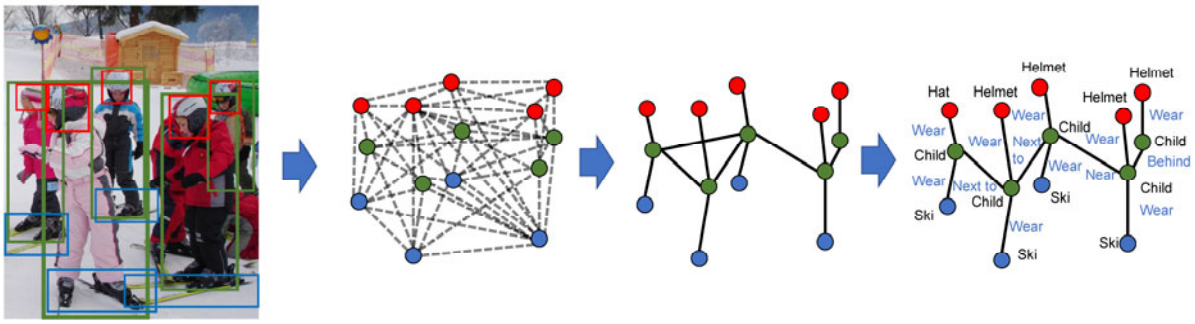
It can be seen that the essence of visual knowledge is based on the reconstruction of computer graphics. It not only provides the possibility of logical reasoning in the traditional knowledge representation, but also bears the characteristics of image perception and image memory in imagery thinking, and therefore is a new form of knowledge representation supporting mental imagery reasoning and visual thinking.

The construction of visual knowledge is a systematic project, which requires interdisciplinary knowledge of machine learning, computer graphics, etc. At present, study on the scene graph (Krishna et al., 2017) is the closest to logical thinking in visual knowledge and further links it with visual objects. The scene graph is a directed graph representing the semantic information of the scene. It explicitly expresses the visual objects in the image and the visual relationship between them.

The scene graph can provide clear reasoning logic for existing deep learning algorithms: First, it converts the visual media (images, videos) into structured data to facilitate the measurement of the model's understanding; second, structured scene graphs also promote the understanding and generation of complex scenes (Zhang HW et al., 2017). Through understanding a large number of scene structures, the existing AI algorithms can achieve the deconstruction of reality, decomposing the scene into more fine-grained components that enable abstract thinking and provide operable and reasoning objects for subsequent creative design. At present, the scene graph has supported a series of applications such as visual description generation (Yang X et al., 2019), visual question answering (Norcliffe-Brown et al., 2018), graph question answering (Hudson and Manning, 2019), visual reasoning (Haurilet et al., 2019), visual matching (Liu et al., 2019), and image generation (Johnson et al., 2018).

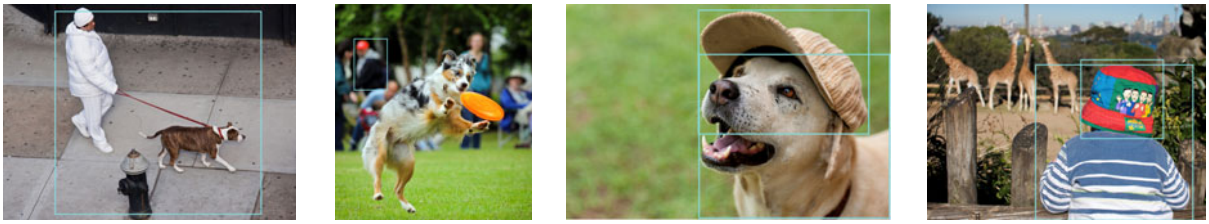
For the construction and deployment of the scene graph, two-stage methods are mainly adopted (Yang JW et al., 2018); that is, the objects are detected first and then the visual relationship is built based on the detected objects. As shown in Fig. 2, the process can be divided into several steps: first, detect the object position; second, reduce the number of plausible visual relationships; finally, classify the objects and the relations. For an image scene graph, the construction difficulty comes mainly from two aspects: (1) There are multiple varying visual relationships between the same subject and object; as shown in Fig. 3, (watch, walk with) are both applicable for the relationship between person and dog; (2) For the same visual relationship, the appearance characteristics of the subject and the object are also very different. As shown in Fig. 3, for the same predicate "wear," the contents of different images are completely different.

Apart from the field of visual reasoning, the deployment of the visual scene graph has also boosted



(a) (b) (c) (d)

Fig. 2 The two-stage method of constructing a scene graph: (a) detecting visual objects as graph nodes; (b) constructing a densely connected graph; (c) pruning the densely connected graph to a sparse graph; (d) determining the relationships between graph nodes



(a) (b) (c) (d)

Fig. 3 Visual relationships in a scene graph: (a) person-walks with-dog; (b) person-watches-dog; (c) dog-wears-hat; (d) child-wears-hat

the quality of image generation because of its deeper understanding of the objects to be created and the relationships among them (Gu et al., 2019; Mittal et al., 2019; Tripathi et al., 2019; Herzig et al., 2020). For example, Johnson et al. (2018) adopted a pipeline which first extracts the features of the scene graph by graph convolutional neural networks, and then predicts the scene layout based on the crucial properties of the visual concepts. This is an explicit measure of projecting the abstract visual knowledge to the image. The results demonstrate that scene graph based methods indeed conform better to relationships of the objects. This is key to respecting the creative ideas generated by the machine.

For video scene graphs, compared with image scene graphs, there are three additional features:

1. The visual relationship changes over time.

2. The temporal information in the video is able to distinguish visual relationships that are difficult to distinguish in the image, such as the difference between a walk and a run.

3. Some of the visual information is present only in the videos.

To address the above difficulties, we adopt counter-factual (Chen et al., 2019) technology (i.e., CMAT) to extract the individual contribution of each local factor in the scene graph generation process, namely, to find important nodes and edges, and try to avoid these important nodes from being misclassified. This allows the overall consistency and local sensitivity of the scene graph to be maintained at the same time. This improves the interpretation and application effect. In the process of video scene graph construction, we propose an iterative graph learning method that gradually learns the graph structure for a video (Shen et al., 2020). These methods have improved the ability and reliability of scene graphs to model visual scenes to a certain extent, and provide a foundation for future research on mental imagery reasoning and visual thinking.

4 Conclusions and future work

The scene graph is a scheme for visual knowledge representation. It provides channels for "machine learning + logical reasoning" and further

provides a basis for the implementation of the idea of visual knowledge. One interesting direction that already emerged is to incorporate the logical graph representations (such as semantic network, knowledge graph, and parsing tree) from other modalities (such as language and audio) into scene graph construction, or to use these graph representations together with the scene graph to improve the performance of downstream computer vision or multimedia tasks such as grounded image captioning (Zhang W et al., 2021), video captioning (Zhang W et al., 2020), and phrase grounding (Mu et al., 2021). At present, visual scene graph is gradually attracting attention in the fields of computer vision, language understanding, and multimedia. Researchers are working on the problems of more fine-grained scene graph construction (Bau et al., 2019; Li YL et al., 2019), more visual interaction between objects (Zareian et al., 2020), better use of external knowledge (Yu et al., 2017; Gu et al., 2019), and how to construct multimedia scene graphs including multi-modal data such as audio and video (Li ML et al., 2020). This showcases the importance of visual knowledge and visual thinking. It is believed that, in the near future, these studies will further guide the deepening of machine creativity.

Contributors

Yueting ZHUANG provided the main idea and outlined the manuscript. Siliang TANG drafted the manuscript. Yueting ZHUANG and Siliang TANG revised and finalized the paper.

Compliance with ethics guidelines

Yueting ZHUANG and Siliang TANG declare that they have no conflict of interest.

References

Arnheim R, 1997. Visual Thinking. University of California Press, San Francisco, USA.

Bau D, Zhu JY, Wulff J, et al., 2019. Seeing what a GAN cannot generate. Proc IEEE/CVF Int Conf on Computer Vision, p.4501-4510. https://doi.org/10.1109/ICCV.2019.00460

Chen L, Zhang HW, Xiao J, et al., 2019. Counterfactual critic multi-agent training for scene graph generation. Proc IEEE/CVF Int Conf on Computer Vision, p.4612-4622. https://doi.org/10.1109/ICCV.2019.00471

Denis M, 1991. Imagery and thinking. In: Cornoldi C, McDaniel MA (Eds.), Imagery and Cognition. Springer, New York, NY, USA, p.103-131. https://doi.org/10.1007/978-1-4684-6407-8_4

Elgammal A, Liu BC, Elhoseiny M, et al., 2017. CAN: creative adversarial networks, generating "art" by learning about styles and deviating from style norms. https://arxiv.org/abs/1706.07068

Gazzaniga MS, 1967. The split brain in man. *Sci Am*, 217(2): 24-29. https://doi.org/10.1038/scientificamerican0867-24

Gu JX, Zhao HD, Lin Z, et al., 2019. Scene graph generation with external knowledge and image reconstruction. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.1969-1978. https://doi.org/10.1109/CVPR.2019.00207

Haurilet M, Roitberg A, Stiefelhagen R, 2019. It's not about the journey; it's about the destination: following soft paths under question-guidance for visual reasoning. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.1930-1939. https://doi.org/10.1109/CVPR.2019.00203

Herzig R, Bar A, Xu HJ, et al., 2020. Learning canonical representations for scene graph to image generation. 16th European Conf on Computer Vision, p.210-227. https://doi.org/10.1007/978-3-030-58574-7_13

Hudson DA, Manning CD, 2019. GQA: a new dataset for real-world visual reasoning and compositional question answering. https://arxiv.org/abs/1902.09506

Johnson J, Gupta A, Li FF, 2018. Image generation from scene graphs. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.1219-1228. https://doi.org/10.1109/CVPR.2018.00133

Kolodner J, 2014. Case-Based Reasoning. Morgan Kaufmann, San Mateo, USA.

Krishna R, Zhu YK, Groth O, et al., 2017. Visual genome: connecting language and vision using crowdsourced dense image annotations. *Int J Comput Vis*, 123(1):32-73. https://doi.org/10.1007/s11263-016-0981-7

Li ML, Zareian A, Zeng Q, et al., 2020. Cross-media structured common space for multimedia event extraction. https://arxiv.org/abs/2005.02472

Li YL, Xu L, Huang XJ, et al., 2019. HAKE: human activity knowledge engine. https://arxiv.org/abs/1904.06539v2

Liu DQ, Zhang HW, Zha ZJ, et al., 2019. Referring expression grounding by marginalizing scene graph likelihood. https://arxiv.org/abs/1906.03561v1

McCarthy J, Minsky ML, Rochester N, et al., 2006. A proposal for the Dartmouth summer research project on artificial intelligence. *AI Mag*, 27(4):12-14.

Mittal G, Agrawal S, Agarwal A, et al., 2019. Interactive image generation using scene graphs. https://arxiv.org/abs/1905.03743

Mu Z, Tang S, Tan J, et al., 2021. Disentangled motif-aware graph learning for phrase grounding. Proc 35th AAAI Conf on Artificial Intelligence.

Norcliffe-Brown W, Vafeais E, Parisot S, 2018. Learning conditioned graph structures for interpretable visual question answering. https://arxiv.org/abs/1806.07243v1

Pan YH, 2019. On visual knowledge. *Front Inform Technol Electron Eng*, 20(8):1021-1025. https://doi.org/10.1631/FITEE.1910001

Pan YH, 2020a. Miniaturized five fundamental issues about visual knowledge. *Front Inform Technol Electron Eng*, online. https://doi.org/10.1631/FITEE.2040000

Pan YH, 2020b. Multiple knowledge representation of artificial intelligence. *Engineering*, 6(3):216-217. https://doi.org/10.1016/j.eng.2019.12.011

Radford A, Metz L, Chintala S, 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. https://arxiv.org/abs/1511.06434

Shen K, Wu LF, Xu FL, et al., 2020. Hierarchical attention based spatial-temporal graph-to-sequence learning for grounded video description. Proc 29th Int Joint Conf on Artificial Intelligence, p.941-947. https://doi.org/10.24963/ijcai.2020/131

Tripathi S, Bhiwandiwalla A, Bastidas A, et al., 2019. Using scene graph context to improve image generation. https://arxiv.org/abs/1901.03762

Yang JW, Lu JS, Lee S, et al., 2018. Graph R-CNN for scene graph generation. Proc 15th European Conf on Computer Vision, p.690-706. https://doi.org/10.1007/978-3-030-01246-5_41

Yang X, Tang KH, Zhang HW, et al., 2019. Auto-encoding scene graphs for image captioning. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.10677-10686. https://doi.org/10.1109/CVPR.2019.01094

Yang XY, Mei T, Xu YQ, et al., 2016. Automatic generation of visual-textual presentation layout. *ACM Trans Multim Comput Commun Appl*, 12(2):33. https://doi.org/10.1145/2818709

Yu RC, Li A, Morariu VI, et al., 2017. Visual relationship detection with internal and external linguistic knowledge distillation. Proc IEEE Int Conf on Computer Vision, p.1068-1076. https://doi.org/10.1109/ICCV.2017.121

Zareian A, Karaman S, Chang SF, 2020. Weakly supervised visual semantic parsing. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.3733-3742. https://doi.org/10.1109/CVPR42600.2020.00379

Zhang HW, Kyaw Z, Chang SF, et al., 2017. Visual translation embedding network for visual relation detection. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.3107-3115. https://doi.org/10.1109/CVPR.2017.331

Zhang W, Wang XE, Tang S, et al., 2020. Relational graph learning for grounded video description generation. Proc 28th ACM Int Conf on Multimedia, p.3807-3828. https://doi.org/10.1145/3394171.3413746

Zhang W, Shi H, Tang S, et al., 2021. Consensus graph representation learning for better grounded image captioning. Proc 35th AAAI Conf on Artificial Intelligence.