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

观点：

视觉知识：智能创意初探

庄越挺‡，汤斯亮

浙江大学计算机科学与技术学院，中国杭州市，310027

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

本文编译自 Zhuang YT, Tang SL, 2021. Visual knowledge: an attempt to explore machine creativity. *Front Inform Technol Electron Eng*, 22(5):619-624. https://doi.org/10.1631/FITEE.2100116

1 导论——从思维科学谈起

长期困扰人工智能（Artificial Intelligence，以下简称 AI）界的一个问题是：AI 能创造吗？或者说，推理能创造吗？为了回答这个问题，本文从思维科学的角度出发，探讨视觉知识表示及其在机器创造中的潜在应用. 具体而言，本文首先列举基于形象思维推理的相关研究，随后介绍一种特殊类型的视觉知识表示——视觉场景图，最后详细回顾视觉场景图的构建问题及其潜在应用. 所有证据表明，视觉知识和视觉思维不仅可以提高当前 AI 任务的性能，而且可以用于机器创造的实践.

AI 已迎来新的发展时代. 现有算法在聚类、分类、逻辑推理和证明方面均取得了不错的效果. 然而，回顾 AI 的早期定义(McCarthy et al., 2006)——“让机器像人类一样认知、思考和学习”，现有 AI 算法仍然存在很大差距，特别是在类人的创造性方面存在严重不足.

早在上世纪 80 年代，上一个 AI 热潮方兴未艾，但学术思想却处于混乱状态. 在这样背景下，我国科学家钱学森先生提出，中国应该创建思维科学(noetic science)来研究思维活动的规律和形式. 思维科学是处理意识与大脑、精神与物质、主观与客观的科学，钱学森主张将发展思维科学同 AI、智能计算机的工作结合起来，并认为思维科学应该从构筑抽象思维、形象（直感）思维、社会思维以及特异思维（灵感思维）等方面入手. 钱老的主张从某些层面上也暗合脑科学在 60 年代的早期研究成果(Gazzaniga, 1967)，即左脑主司语言、逻辑分析、推理、抽象、计算语言记忆、书写等逻辑思维；右脑主司直觉、情感、图形知觉、形象记忆、美术、音乐节奏、舞蹈想象、视觉、知觉身体协调、灵感等形象思维. 他的这些想法与建议突破了当时 AI 逻辑思维的主流框架，为智能创意的实现提供了思路. 时至今日，这些想法与建议仍具有非常重要的指导意义与理论价值.

近年来，随着 AI 技术的演进，主流学术界如 *Science*、*Nature* 等顶级学术期刊，AAAI、IJCAI 等顶级 AI 会议也开始大力关注创造性智能相关研究. 其核心问题是对创造性思维的模拟，即 AI 能创造吗？推理能创造吗？

以平面广告设计这种创意行为为例，其中涉及对象形状、空间关系、色彩、纹理等大量视觉信息，人类设计师需要善于在信息不完备条件下进行形象思维指导下的推理，这种推理是一种跳跃性、非连续的思维过程. 在这个思维过程中，人类会利用“心象”(mental imagery) (Denis, 1991)这一工具，即在大脑中排列、组合、重建、操作相关的视觉信息，来探索、想象、推理符合要求的设计方案. 这一过程也被称为“视觉思考”(visual thinking) (Arnheim, 1997). 为了让机器具备推理和创造的能力，恰当地保存视觉知识至关重要. 原因在于，视觉知识是让算法理解视觉世界的基础. 现

‡通讯作者

实世界中，表示常识和对象之间关系的方式是机器实现创造的第一步. 目前的 AI 算法在创造性思维方面已经取得一些进展，但在上述“心象”推理与“视觉思考”方面仍然值得深入探索.

2 形象思维推理相关研究

可支撑形象思维推理的相关研究最早可以追溯到上世纪 80 年代的基于实例的推理(case-based reasoning, CBR) (Kolodner, 2014). CBR 是 AI 和认知科学的典型范式，是一种基于类比的推理方法. CBR 的基本思想是基于记忆（范例）来模拟推理的过程. 它的基本步骤包括：

1）提取：针对给定目标问题，从记忆库中检索出与求解问题相关的案例；

2）再用：将针对先前案例的问题求解方案，映射到目标问题中；

3）修改：在真实世界（或仿真）中测试新的解决方案，如有需要，修改之；

4）保存：在解决方案成功用于目标问题后，将此新经验以新案例的形式存储在数据库中.

CBR 常常用于机械师修理、医生诊治、法官断案等推理系统中. 以广告创意设计为例：设有广告范例 $C_1, C_2, \ldots, C_m$. $g(C_i, P_i)$表示从范例 C_i 中获取特性 P_i. 广告画面的视觉特征可包括：广告的渲染、广告标语、色彩搭配、布局、效果处理等. 故 C_{new} 的最终结果可以描述为:

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

其中，C_{new} 是当前的新设计，$\tilde{\cap}$ 表示组合的广义运算，C_i 对 C_{new} 的贡献与 $g(C_i, P_i)/C_{\text{new}}$ 成正比. 可以发现 m 越大，C_i 和 C_{new} 就越不相似.

在广告创意设计的 CBR 系统中，一个广告设计可以被抽象成一种由“视觉”（视觉特征）与“符号”（位置、组合方式）共同组成的推理系统，在一定程度上同时考虑了形象思维与逻辑思维. 尽管这个思路起源于 80 年代，但我们可以看到 CBR 仍然影响着近期研究，如 2017 年的 *ACM Trans Multim Comput Commun Appl* 的最佳论文(Yang XY et al., 2016).

近年来，对抗生成网络(generative adversarial networks，GANs) (Radford et al., 2015)在图像生成方面取得很大进展. 通过判别网络与生成网络之间的零和博弈，GAN 使生成的样本分布不断拟合真实的样本分布，借此方法获得的图像生成器可以产生以假乱真的图像. 其中创意对抗网络(creative adversarial networks，CANs) (Elgammal et al., 2017)对 GAN 稍作改造，加入风格判断，产生的创意画作通过了图灵测试（见图 1）.

图 1 创意对抗网络生成的创意画作(Elgammal et al., 2017)

尽管目前 GAN 以及其变种为智能创意方面带来可喜进展，但此类方法仍然存在很多问题，如 GAN 容易出现坍缩(mode collapse)，也很容易忽略一些对象(mode drop) (Bau et al., 2019). 究其原因，GAN 本质上是分布拟合，逻辑思维以及形象思维的缺失使其无法进行原始创新.

3 视觉知识表达与视觉场景图

GAN 在生成过程中的一系列问题使我们意识到真正类人的智能创意需要逻辑思维与形象思维的有效协同. 如何在一个框架下协同两种截然不同的思维方式是当下智能创意亟需解决的关键核心问题. 近年，潘云鹤院士提出视觉知识表达(Pan, 2019, 2020a)和多重知识表达(Pan, 2020b)理论，系统阐述了“视觉知识”这种可以有效融合逻辑思维与形象思维的新型知识表达. 他认为视觉知识具有以下特征：

1）能表达对象的空间形状、大小和关联关系，以及色彩和纹理；

2）能表达对象的动作、速度及时间关系；

3）能进行对象的时空变换、操作与推理，包括形状变换、动作变换、速度变换、场景变换、各种时空类比、联想和基于时空推理结果预测等.

由此可知，视觉知识的本质是基于计算机图形学的重构，它既提供传统知识表达中逻辑推理的可能，也具备形象思维中图形知觉、形象记忆的特点，是一种可以支撑“心象”推理与“视觉思考”的新知识表达形式.

视觉知识的构建是一个系统工程，需要对机器学习、计算机图形学等多个学科知识进行跨学科整合. 目前最接近视觉知识中的逻辑思维、并将其与视觉对象链接的工作是“视觉场景图”(scene graph) (Krishna et al., 2017). 视觉场景图是一种表示场景语义信息的有向图，可以显式地表达图像中所包含的视觉对象以及对象之间的视觉关系.

视觉场景图可以为现有深度学习算法提供清晰的推理逻辑：一方面，视觉场景图将可视媒体（图像、视频）转化成结构化数据，以便于衡量模型的理解能力；另一方面，结构化的场景图也促进复杂场景的理解与生成(Zhang HW et al., 2017). 通过对大量场景的结构化理解，有助于帮助目前的 AI 算法实现对现实的解构，将场景分解到可进行抽象思维的更细粒度，为后续的创意设计提供可操作、推理的对象、对象范畴与属性. 目前，视觉场景图已支撑视觉描述生成(Yang X et al., 2019)、视觉问答(Norcliffe-Brown et al., 2018)、图问答(Hudson and Manning, 2019)、视觉推理(Haurilet et al., 2019)、视觉匹配(Liu et al., 2019)、图像生成(Johnson et al., 2018)等一系列应用任务.

目前视觉场景图构建的技术路线主要采用两阶段方法(Yang JW et al., 2018)，即先检测物体框，再检测视觉关系.具体构建过程如图 2 所示：首先检测物体位置，然后减少视觉关系组合，最后对物体和视觉关系分类. 对于图像场景图而言，其构建难点主要有两方面：（1）同样的主语和宾语之间存在多种不同的视觉关系，如图 3 所示：“person”和“dog”之间存在（“watch”和“walk with”）等多种不同的视觉关系；（2）在同一种视觉关系中，主语和宾语的外表特征也存在非常大的差异，如图 3 中对于同一个视觉关系“wear”，不同图像的内容也完全不同.

除视觉推理领域之外，视觉场景图的部署还提高了图像生成的质量，因为它对要创建的对象及其之间的关系有更深入理解(Gu et al., 2019; Mittal et al., 2019; Tripathi et al., 2019; Herzig et al., 2020). 例如，Johnson 等(2018)采用一种流程，首先通过图卷积神经网络提取场景图特征，然后基于视觉概念的关键属性来预测场景布局. 这是将抽象视觉知识投射到图像上的一种明确措施. 结果表明，基于场景图的方法更符合对象之间的关系. 这也是重视机器产生的创造性想法的关键.

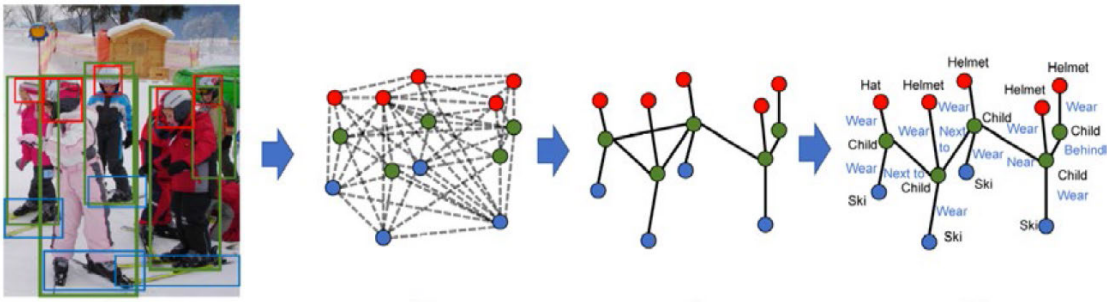

图 2　两阶段方法构建视觉场景图：（a）将视觉对象检测为图节点；（b）构造一个紧密连接的图；（c）将紧密连接的图修剪为稀疏图；（d）确定图节点之间的关系

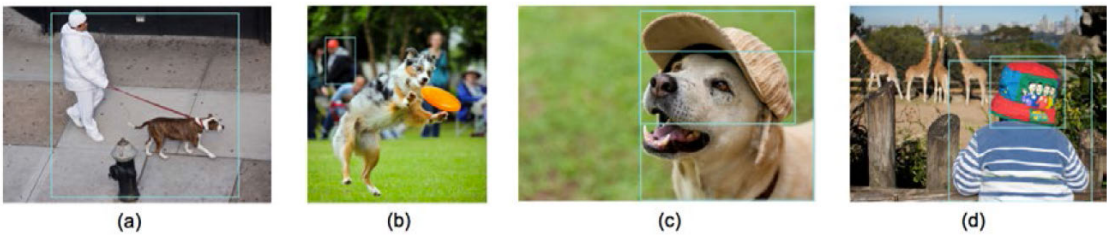

图 3　图像场景图中的视觉关系：（a）人-遛-狗；（b）人-看-狗；（c）狗-戴-帽子；（d）孩子-戴-帽子

与图像场景图相比，视频场景图有 3 个额外特点：

1）视觉关系随着时间变化；

2）视频中的时序信息能够区分难以通过图像区分的视觉关系，如行走和跑步之间的差异；

3）部分视觉信息只在视频中出现.

针对上述困难，在图像场景图构建方面，我们采用反事实(Chen et al., 2019)技术，利用反事实基准模型，分离出场景图生成过程中每个局部预测的贡献，即发现重要的节点与边，并尽量避免这些重要的节点被错分. 这使场景图的整体一致性与局部敏感性得以同时保持，提升了最终构图的解释性与应用效果. 在视频场景图构建过程中，我们提出一种与初始状态无关的迭代图学习方法(Shen et al., 2020)，建立一种与下游任务相关但与初始构图方法无关的迭代图优化方法，通过多次迭代来构建视频场景图. 这些方法从某种程度上提升了场景图在视觉场景建模上的能力与可靠性，为后续“心象”推理与“视觉思考”研究提供了可操作的基础.

4 总结与展望

视觉场景图是视觉知识的一种表达，它可以为“机器学习+逻辑推理”提供渠道，并进一步为视觉知识思想的实施提供基础. 目前一个有趣的方向是将源于其他形式（如语言和音频）的逻辑图表示（如语义网络、知识图谱和解析树）结合到场景图的构建中，或者将这些图表示与场景图一起使用，以提高下游计算机视觉或多媒体任务的性能，例如图像对齐字幕生成(Zhang W et al., 2021)、视频字幕生成(Zhang W et al., 2020)和短语到图片区域的对齐(Mu et al., 2021). 目前视觉场景图相关研究逐渐引起计算机视觉、语言理解和跨媒体领域的关注，人们正在关注更细粒度的场景图构建(Bau et al., 2019; Li YL et al., 2019)、更多物体间的视觉关系交互(Zareian et al., 2020)、更好的外部知识利用(Yu et al., 2017; Gu et al., 2019)以及包括音视频等多模态数据的跨媒体场景图的构建(Li ML et al., 2020). 这体现了视觉知识和视觉思维的重要性. 相信在不久的将来，这些研究工作也将进一步引领智能创意技术的深入，并在智能创意中发挥重要作用.

作者贡献

庄越挺提供主要思想及概述. 汤斯亮起草论文手稿. 二人共同完成修改和定稿.

遵守道德准则声明

两位作者声明本文不存在利益冲突.

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.