受脑启发的机器人认知抓取决策模型

左国玉[1,2], 刘洪星[1,2], 龚道雄[1,2], 阮晓钢[1,2]
(1. 北京工业大学信息学部, 北京 100124; 2. 北京市计算智能与智能系统重点实验室, 北京 100124)

摘 要: 为了让机器人获得更加通用的能力,抓取是机器人必要掌握的技能. 针对目前大多数机器人抓取决策方法存在物品特征理解浅显,缺乏抓取先验知识,导致任务兼容性较差的问题,同时受大脑中分区分块功能结构的启发,提出了将物品感知、先验知识和抓取任务融合的认知决策模型. 该模型包含卷积感知网络、记忆图网络和贝叶斯决策网络三部分,分别实现了物品能供性(affordance)提取、抓取先验知识推理和联想,以及信息融合编码决策,三部分之间的信息流以语义向量的形式传递. 利用 UMD part affordance 数据集、该文构建的抓取常识图和决策数据集对3个网络分别进行训练,认知决策模型的测试准确率达到99.8%,并且抓取位置可视化结果展示了决策的正确性. 该模型还能判断物品是否属于当前任务场景,以决策是否抓取以及选择什么部位抓取物品,有助于提高机器人实际场景的应用能力.

关键词: 机器人抓取; 认知模型; 决策模型; 物品感知; 记忆图; 脑启发
中图分类号: U 461; TP 308 **文献标志码**: A **文章编号**: 0254 - 0037(2021)08 - 0863 - 11
doi: 10.11936/bjutxb2020120034

Brain-inspired Decision-making Model for Robot Cognitive Grasping

ZUO Guoyu[1,2], LIU Hongxing[1,2], GONG Daoxiong[1,2], RUAN Xiaogang[1,2]
(1. Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China;
2. Beijing Key Laboratory of Computing Intelligence and Intelligent Systems, Beijing 100124, China)

Abstract: To obtain general purpose ability in human's life and work, robots first need to master the skill of grasping objects. However, most current robot grasping decision-making methods have many problems such as simple understanding of object features, lack of grasping prior knowledge and poor task compatibility. Inspired by the functional structure of partitions and blocks in the brain, this paper proposed a decision model that integrates object perception, prior knowledge and grasping task. The model consists of three parts: convolutional perception network, memory graph network and Bayesian decision network, which realize the functional affordance extraction of objects, grasping prior knowledge reasoning and association, and decision-making with information fusion, respectively. Three networks were respectively trained on the UMD part affordance dataset, self-built common-sense graph, and self-built decision dataset. Test on the cognitive model verified its good performance with the accuracy of 99.8%. Results show that it can make reasonable decisions, including the ability whether the object belongs to the current task scene and the ability whether and where to grasp, which can help improve the

收稿日期: 2020-12-31
基金项目: 国家重点研发计划资助项目(2018YFB1307004);国家自然科学基金资助项目(61873008);北京市自然科学基金资助项目(4182008, 4192010)
作者简介: 左国玉(1971—), 男, 教授, 主要从事机器人控制、机器人学习、智能计算方面的研究, E-mail: zuoguoyu@bjut.edu.cn

robot's availability in real applications.

Key words: robot grasp; cognitive model; decision model; object perception; memory graph; brain inspiration

随着机器人技术的进步,机器人正在代替人类完成一些重复、简单的操作. 然而,为了让机器人获得更加通用的能力,抓取技能是机器人必须要掌握的. 抓取是人类行为中常见但复杂的综合性行为,其整合了感知、认知决策和动作执行以及其间的协调与配合,体现了人类的认知能力和操纵能力. 研究者们在机器人智能抓取领域已经取得了一些进展. 一些工作[1-3]将机器人抓取检测看作计算机视觉问题,并使用深度学习方法以目标检测的方式进行研究. 这些深度神经网络结构依赖于卷积神经网络(convolutional neural network, CNN)[4]. CNN 是受哺乳动物的视觉通路启发而产生的,并且在空间和特征处理方面有很好的表现. 机器人利用深度神经网络赋予的视觉感知能力对抓取位置进行回归或分类,其中抓取位置具体指示了机器人末端执行器以怎样的姿态抓起物体. 然而,目标检测的方法不能满足机器人对物品更深层次的探索和理解. 因此相关学者对 affordance 检测展开了研究. Affordance 检测和目标检测最大的区别在于关注的物品特征形式不同. Affordance 检测关注的是物品与环境的交互特征. Affordance 是指用一个物体进行不同行为的可能性,这个概念最早是由心理学家吉布森[5]提出. Affordance 的概念用于描述物品的功能特性,在机器人抓取和操作的研究中得到了广泛的应用[6-8]. 一些工作[9-10]借助深度学习方法,使用视觉输入学习 affordance 表征,其中 affordance 由图像中物品的具有特征区分性的部分表示. 物品的抓取方式与物品的 affordance 密切相关. Kokic 等[11]利用 CNN 在点云上对 affordance 进行编码和检测并使用 affordance 来建模任务、对象和抓取动作之间的关系. 类似地,Chu 等[12]表明基于部分的物品表征有利于 affordance 检测,因为一些物品部分分别具有独特的特征但与其他物品又具有共性,所以可以推广到新颖的物品上使用. Zeng 等[13]使用 CNN 将视觉观察(例如图像)映射到感知的 affordance 上用以关联物品和动作. 在物品感知中,affordance 检测使得机器人可以获取物品与环境的交互特征,并使得物品特征以更加基元化、更加普遍的形式表现,为机器人的抓取操作提供了重要的信息. 然而,这些模型没有考虑抓取相关的约束条件(例如任务),也没有使用先验知识指导机器人最终的抓取决策. 值得注意的是,视觉感知的作用更像是一个环境感受传感器,机器人并不能只依靠传感器实现完整的推理和决策,这最终会导致不灵活、非鲁棒的抓取表现.

物品感知在一定程度上实现了物品的分割和解析,并且感知结果会在抓取决策阶段作为影响因素被考虑. 目前抓取决策的方法可分为基于概率逻辑的方法和基于学习的方法. Ardón 等[14]为了得到物品抓取 affordance 的概率分布,利用马尔可夫逻辑网络建立了知识图表征. Antanas 等[15]使用概率逻辑模块,通过利用物品部分的语义、物品的属性和任务约束来提高抓取能力. Fang 等[16]提出了一种面向任务的抓取网络,用于联合预测面向任务的抓取和后续操作动作. 在基于学习的方法研究中,Karaoguz 等[17]对抓取矩形建议网络检测到的抓取矩形按照得分进行排序,以得分最高的抓取矩形作为目标抓取位置. Kasaei 等[18]通过人机交互的方式学习抓取,示教者使用示教的方式向机器人演示一个物体的可能的抓取方式. 这些方法中,概率逻辑规则使抓取决策过程具有可解释性. 然而,手工设计的逻辑规则的设计和学习通常是复杂的. 视觉输入的深度学习方法是黑箱学习. 虽然该方法避免了手工规则设计,但可解释性较低.

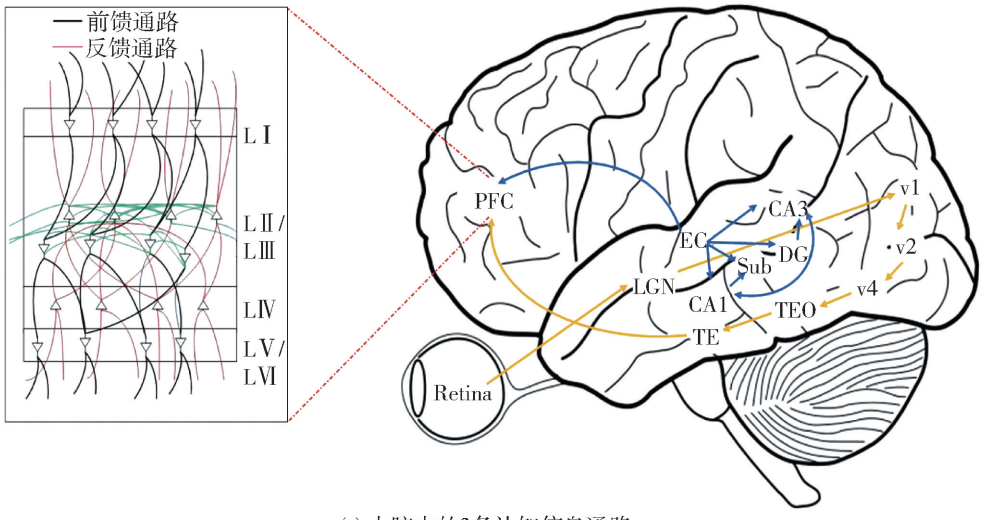
抓取行为本质上是大脑综合认知的一种外部表现,若只考虑利用一方面的能力来实现智能抓取是很困难的. 因此抓取模型应该被赋予多种类似人一样的认知功能. 不可否认,在机械任务层面机器人和生物的抓取表现是很相似的. 然而,目前机器人和人类对抓取的认知在决策层面上还有很大的差距,而且这种决策能力会直接影响机器人后续抓取动作的执行和操作. 对于机器人而言如何将人类认知中形而上的功能(例如记忆、视觉感知和大脑皮层推理)整合到一起是必要且亟待解决的问题.

人类大脑集合了多种类型的认知功能,受人类大脑分区分块的功能结构的启发,本文提出了一种认知抓取决策模型. 模型包含了 3 个信息通路:1)受视觉腹部通路功能启发构建了一个卷积神经网络以实现物品空间信息和特征信息的提取;2)受海马体信息通路功能启发构建了一个图神经网络以实现

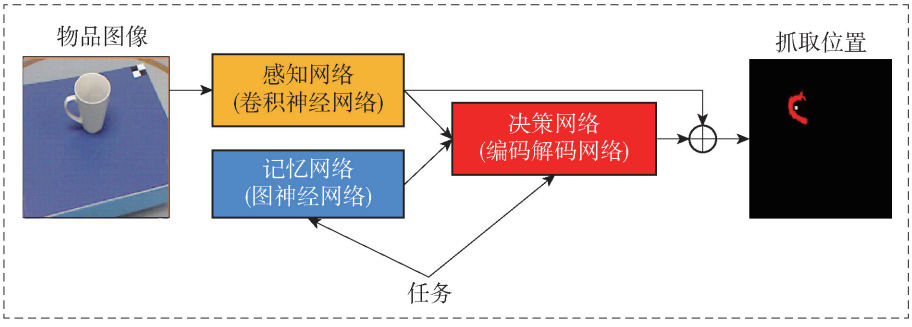
数据的存储以及推理检索;3)受皮质柱信息通路功能启发构建了一个贝叶斯编码解码网络以实现信息的融合和最终的决策. 因此通过模仿人类大脑中存在的功能性结构,构建该模型以实现更符合实际应用场景的合理抓取决策.

1　模型结构

生物大脑因其出色地整合了数百种认知功能而在认知方面具有权威性. 视觉和记忆在大脑的认知决策中都起着至关重要的作用. 本文以控制二指机械手抓取为例,提出了一种受脑启发的认知决策模型,以实现合理、灵活的机器人抓取动作决策. 如图 1(a)所示,该模型包含 3 条认知信息通路:负责视觉感知的视觉腹部通路,负责记忆推理和检索的海马体信息通路和负责决策的皮质柱信息通路. 图 1(b)展示了所构建模型整体信息流的传递. 本文采用了 3 种网络架构来分别实现上述 3 种信息加工和信息流的传输功能.



(a) 大脑中的3条认知信息通路



(b) 模型整体信息传递框架

图 1　大脑与认知模型中的信息流

Fig. 1　Information flow in the brain and the cognitive model

1.1　卷积感知网络

在认知视觉信息通路中,原始视觉信息从视网膜外侧膝状核,V1 ~ V4,经一系列连续处理,直到形成复杂的物体表征[19]. 视觉腹侧信息通路通常被认为是识别和处理与形状和颜色相关信息的部分[20-21]. 此外,一些生物抓取行为的研究表明,大脑倾向于将物体形状编码为非整体的、基于部分的格式[22]. 心理学和神经科学都表明,affordance 与抓取行为有着密不可分的联系[23-24]. 因此,本文构建了一个感知网络模拟腹侧信息通路以 affordance 的形式编码视觉信息.

如图 2 所示,感知网络对物品图像进行卷积操作,分割出物品的 affordance 并输出对应类型. 该网络以卷积层为基本结构. 利用预先训练好的 5 个卷积块作为第 1 个编码块来提取目标图像的低层特征. 然后,采用 4 个反卷积层[25]作为第 2 个编码块进行高层特征编码. 图像中可区分的低层特征(低分辨率)通过第 1 个编码块学习,然后将这些特征语义编码到像素空间(高分辨率)获取图像中物体的 affordance 分类. 为了恢复网络提取低层特征时

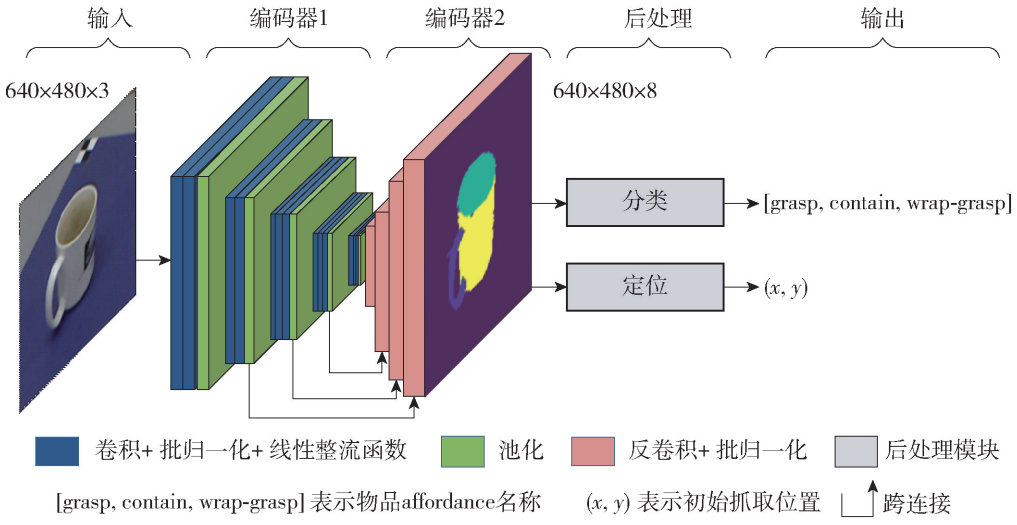


图 2 感知网络结构

Fig. 2 Perception network structure

丢失的空间信息,在高级特征编码过程中,利用跨连接融合第一个编码块不同阶段的空间信息来细化物品的 affordance 分布. 本文采用了 4 个跨连接对 4 种不同分辨率的空间信息进行融合. 为了将感知网络的结果转化为决策网络的可识别输入,使用不需要训练的后处理块提取出 affordance 的语义和像素坐标.

1.2 记忆图网络

海马体与记忆密切相关,海马体信息通路传递着各种与记忆相关的信息. 海马体对于情景记忆的关键作用已经被神经心理学、动物模型、计算模型和人类神经成像[26-28]研究明确地确立了. 计算模型表明,在接收到部分记忆线索后,海马体中的神经元会协调皮层目标部位相关记忆的恢复[29]. 因此,受到海马体神经元之间图形连接信息通路和检索记忆的功能的启发,建立了一个图神经网络作为记忆网络,实现记忆先验的搜索和推理.

一些与图相关的符号如下:定义了一个有向图,$\mathscr{G}=(\mathscr{V},\mathscr{E},\mathscr{R})$. 式中 $\mathscr{V}$、$\mathscr{E}$ 和 $\mathscr{R}$ 分别表示节点的集合、边的集以及关系的集合. 设 $v_i\in\mathscr{V}$ 表示一个节点,$(v_i,r,v_j)\in\mathscr{E}$ 表示一条从 v_i 指向 v_j 的边,其关系为 $r\in\mathscr{R}$. 在常识知识图中许多关系是普遍有效的,被认为是人类的常识. 然而,对于机器人来说,这些关系很难理解和应用. 为了利用有价值的常识记忆作为先验信息,使用一个称为记忆网络的图神经网络来学习常识图. 记忆网络是基于一种图编码器模型:关系图卷积网络(relational graph convolutional network, r-GCN)[30]建立的. 输入线索的触发下,利用图中的关系和节点,对已存储的记忆信息进行推理和搜索,并输出相关结果. 在记忆网络中,使用 r-GCN 层来嵌入图中事实的实体(节点)和关系(边)(例如,三元组(drink, need, contain)). 在记忆网络中,节点和关系用词向量表示. 嵌入过程以关系学习的过程为例. 在局部图邻域中进行操作. 在网络训练中,使用了消息传递框架

$$h_i^{(l+1)}=\sigma\left(\sum_{r\in\mathscr{R}}\sum_{j\in\mathscr{V}_i^r}\frac{1}{c_{i,r}}\boldsymbol{W}_r^{(l)}h_j^{(l)}+\boldsymbol{W}_0^{(l)}h_j^{(l)}\right)\quad(1)$$

式中:$h_i^{(l)}\in\mathbb{R}^{d(l)}$ 为节点 v_i 在神经网络第 l 层的隐藏状态;$d(l)$ 为该层节点表征的维度;V_i^r 表示节点 v_i 在关系 $r\in\mathscr{R}$ 下的邻居索引集合;$c_{i,r}$ 为一个与某个特定关系的归一化常数. $\boldsymbol{W}$ 表示一个权重矩阵,其中 $\boldsymbol{W}_r$ 表示与关系 $r\in\mathscr{R}$ 相关的矩阵,$\boldsymbol{W}_0$ 表示一个自环权值矩阵. 相邻节点的特征向量通过归一化线性变换聚合,并通过元素级激活函数(例如 ReLu(·))传递. 如图 3 所示的记忆网络,对于一个节点(粉红色),按照关系类型聚合相邻节点(白色),并且不同的关系被赋予不同的权重. 利用该框架学习节点间非单一逻辑的信息传输,使神经元间的信息传输可区分. 信息传递过程中的可解释性随着关系特定的传递增加而增加(例如关系的方向和类型). 在训练记忆网络的时候,使用了 DistMult[31]作为得分函数去重建图中的关系,有

$$f(\boldsymbol{v}_i,r,\boldsymbol{v}_j)=\boldsymbol{v}_i^{\mathrm{T}}r\boldsymbol{v}_j\quad(2)$$

经上述处理,记忆网络可以理解节点和关系表示的常识图,并在接受到部分记忆线索之后能检索相关的记忆. 与直接使用知识库进行查询的方法相比,记忆网络使用了消息传递框架能有效地推理和

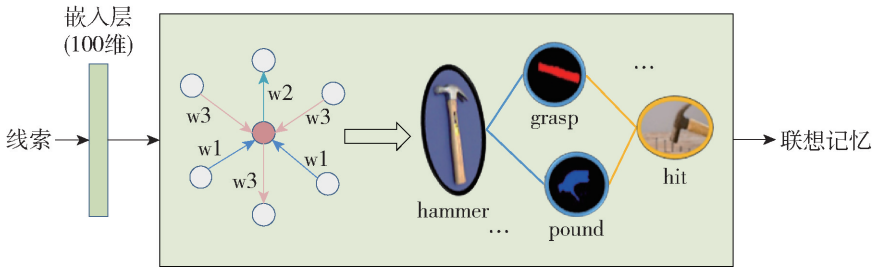


图 3　记忆网络处理信息过程

Fig. 3　Process of information processing by the memory network

学习记忆中的信息，使得记忆检索边的具有逻辑，更加准确.

1.3　贝叶斯决策网络

皮质柱是大脑动力学和皮质信息处理的重要决定因素[32]. 作为感觉处理或运动输出的基本功能单元，皮层柱在皮层的学习和发育中起着重要作用. 6 层细胞构成皮层柱的垂直方向. 皮质柱的每一层都包含不同的细胞类型，并在水平层上通过突触连接[33]. 本文假设皮质柱中信息处理或是一个编码和解码的过程，它会产生一些潜在的特征表达或决策.

本文试图研究和模拟人类潜在的决策过程，以执行完备的抓取动作. 模仿人们的思维方式，将记忆作为先验信息，视觉感知作为观察信息，与任务相关的信息作为约束，帮助机器人实现合理决策. 值得注意的是，人类的行为是由大脑中产生的任务驱动的，因而行动是有目的的. 因此，在决策模型中加入与任务相关的约束是有必要的. 该决策方法符合贝叶斯理论的思想. 故本文基于贝叶斯理论建立了决策网络.

CVAE[34]方法将高维输出空间的分布建模为以输入观测为条件的生成模型，受该方法的启发本文使用了一个条件编码解码去实现决策. 定义 y 表示抓取的决策结果. 决策网络的目标是在给定观测信息 x、先验信息 m 和任务约束 t 的情况下，使 y 的条件对数似然最大化. 网络的条件生成过程如图 4 所示. 高斯隐变量 z 被编码并从先验分布 $p_\theta(z|x,m,t)$ 中进行采样. 输出 y 被解码并从分布 $p_\theta(y|x,z,m,t)$ 中生成. 直观地说，隐变量 z 允许网络对输出 y 的多个条件分布建模，这些条件分布代表可供抓取的潜在选择. 然而，难以处理的隐变量 z 的边缘化问题，使得决策网络的参数估计具有挑战性. 本文使用随机梯度变分贝叶斯框架[35]来解决这个问题. 在 SGVB 中，对数似然的变分下界被用作替代目标函数. 模型的变分下界为

$$\log(p_\theta(y|x,m,t)) \geqslant -D_{KL}(q_\phi(z|x,y,m,t) \parallel p_\theta(z|x,m,t)) + \mathbb{E}_{q_\phi(z|x,y,m,t)} \log(p_\theta(y|x,z,m,t)) \tag{3}$$

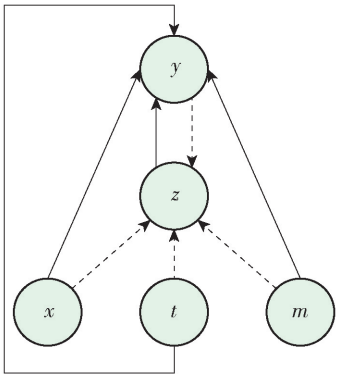


图 4　决策网络图模型

Fig. 4　Graphical model of the decision network

模型的经验目标为

$$\mathscr{L}(x,y,m,t;\theta,\phi) = -D_{KL}(q_\phi(z|x,y,m,t) \parallel p_\theta(z|x,m,t)) + \frac{1}{L}\sum_{l=1}^{L} \log(p_\theta(y|x,z^{(l)},m,t)) \tag{4}$$

式中：$q_\phi(z|x,y,m,t)$ 为识别网络用于估计真实的后验分布 $p_\theta(z|x,y,m,t)$，真实的后验分布 $p_\theta(z|x,y,m,t)$ 表示当给定物品观测信息 x、记忆 m、任务 t 和标签 y 时产生的潜在抓取分布；$p_\theta(z|x,m,t)$ 在这里表示一个条件高斯隐变量 z 的条件先验网络；$p_\theta(y|x,z^{(l)},m,t)$ 表示一个生成网络，$z^{(l)} = g(x,y,m,t,\epsilon^{(l)})$，$\epsilon \in \mathcal{N}(0,I)$，$g(\cdot)$ 是一个使用了重参数化技巧[47]的可微函数；L 表示样本数量.

在模型中，使用了多层感知机去建模识别网络、先验网络和生成网络. 模型有与皮质柱一样的 6 层结构. 训练时的网络结构如图 5 所示，在训练网络时，先验网络和识别网络分别得到的隐变量 z，使用 KL 散度进行处理，目的是使得先验网络逼近识别网络.

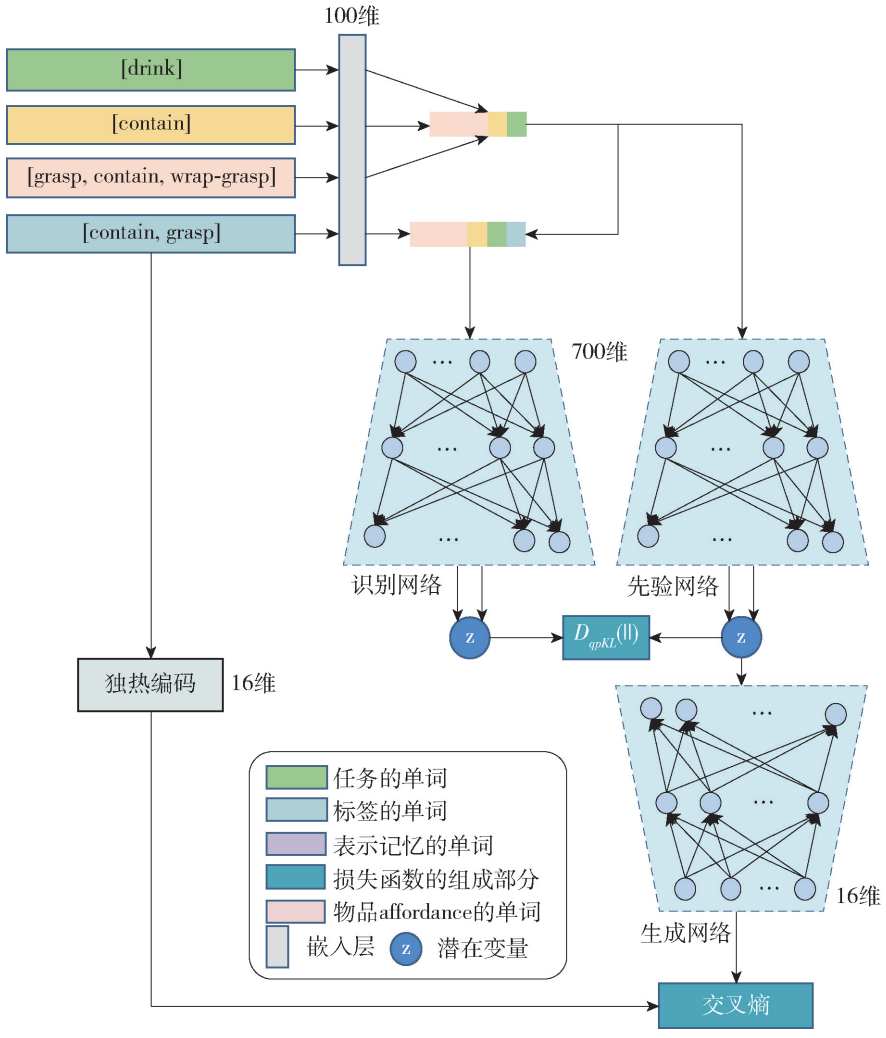


图 5 用于训练的决策网络结构

Fig. 5 Structure of the decision network for training

2 实验结果

本文关注的是给定操作任务时对象的可行性抓取,因此测试以下三方面能力是至关重要的:1) 感知网络的 affordance 检测准确率;2) 记忆网络的记忆联想能力; 3) 决策网络的决策能力.

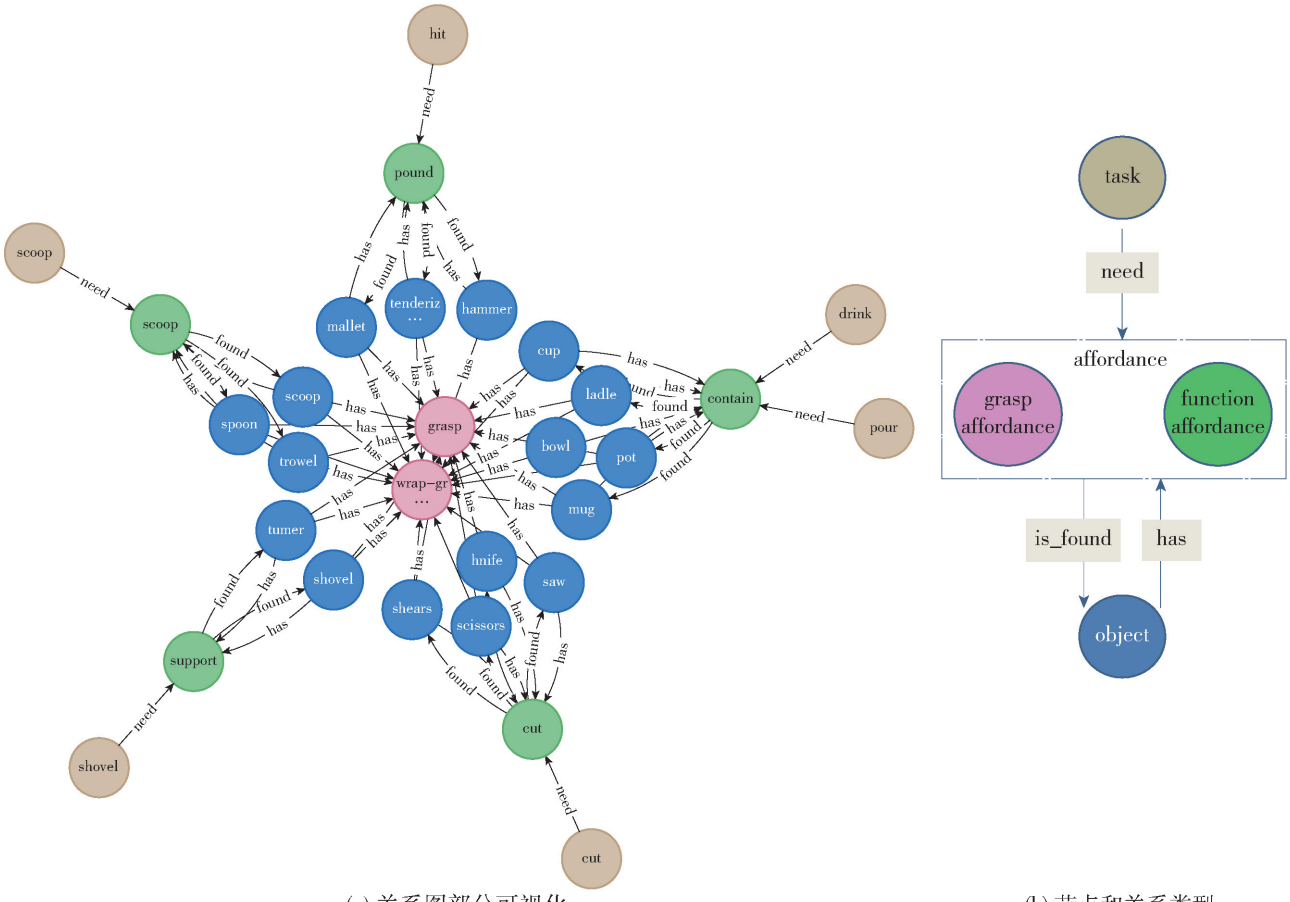
2.1 数据集

基于 Myers 等[36]建立的 UMD part affordance 数据集对认知模型进行了评估. 此数据集包含不同视角的 105 个工具的 RGB-D 图像,并提供了像素级 affordance 标签. 这些工具共有 17 类,包含了 7 类 affordances:grasp、cut、scoop、contain、pound、support 和 wrap-grasp(如表 1 所示). 模型中的感知网络直接对 UMD part affordance 数据集进行处理. 对于记忆网络,需要一个与任务,affordance 和物品相关的抓取常识图作为记忆数据. 但是,目前没有专门用于抓取相关的常识图,或者有类似的图结构数据但是其中包含了大量与本文研究无关的数据,导致无法有效地提取相关数据. 因此,本文使用 Neo4j 图形平台建立了一个抓取的常识图,其关系如图 6 所示. 图 6 中有 140 个节点、315 个关系,包含 3 种类

表 1 工具的 7 种 affordances 描述

Table 1 Description of the seven affordances of tools

affordance	描述
grasp	可以用手围起来进行操作
cut	用于分离出另一个物体(刀刃)
scoop	具有弯曲表面的口可收集柔软材料(勺子)
contain	有很深的腔体来容纳液体(碗的内部)
pound	用于敲打其他物体(锤头)
support	可以容纳松散材料的扁平部分(铲)
wrap-grasp	可以用手和手掌握住(杯子的外部)



(a) 关系图部分可视化　　(b) 节点和关系类型

图 6　抓取常识图

Fig. 6　Common-sense graph for grasping

型的节点：任务节点、affordance 节点和物品节点. 节点之间的关系包括 3 种类型：need、found 和 has.

对于决策网络，本文创建了一个决策数据集. 数据集有 4 个部分：观察到的 affordance 记忆数据、任务和标签. 观察到的 affordance 是从 UMD part affordance 数据集收集的，记忆数据和任务数据是使用建立的抓取常识图进行创建的. 数据集中的每个样本设计为包含观察到的 affordance、任务、记忆的形式，并以单词的形式存储，如表 2 所示. 数据集中有 304 326 个样本. 在决策网络中，使用嵌入层将单词转换为向量.

2.2　Affordance 检测结果

本文在 UMD part affordance 数据集上评估 affordance 检测的表现. 为了进行对比，将 Myers 等[36]和 Sawatzky 等[37]的结果作为基线进行比较. 使用交并比(intersection over union，IoU)作为评价指标来评价 affordance 检测的准确性. 如图 7 所示，本文的方法实现了更高的平均检测精度，在平均 IoU 方面比基于 resnet 的网络高出 14%. 在每类 affordance 的检测中，感知网络也取得了最高的 IoU 值. 这表明，卷积下采样编码和反卷积上采样编码相结合的算法在 UMD part affordance 数据集的 affordance 检测任务上表现很好. 因此，感知网络对物品实现了以 affordance 为基元的物品分解，并且这种以 affordance 形式对物品实现原语理解便于后续

表 2　决策数据集中样本的组成部分

Table 2　Compositions of some samples in the decision dataset

样本	affordances (一个物体)	任务	记忆	标签
1	[contain, grasp, wrap grasp]	[pour]	[contain]	[contain, grasp]
2	[contain, grasp, wrap grasp]	[cut]	[cut]	[cut, none]
3	[cut, grasp]	[pour]	[contain]	[contain, none]
4	[cut, grasp]	[cut]	[cut]	[cut, grasp]
⋮	⋮	⋮	⋮	⋮

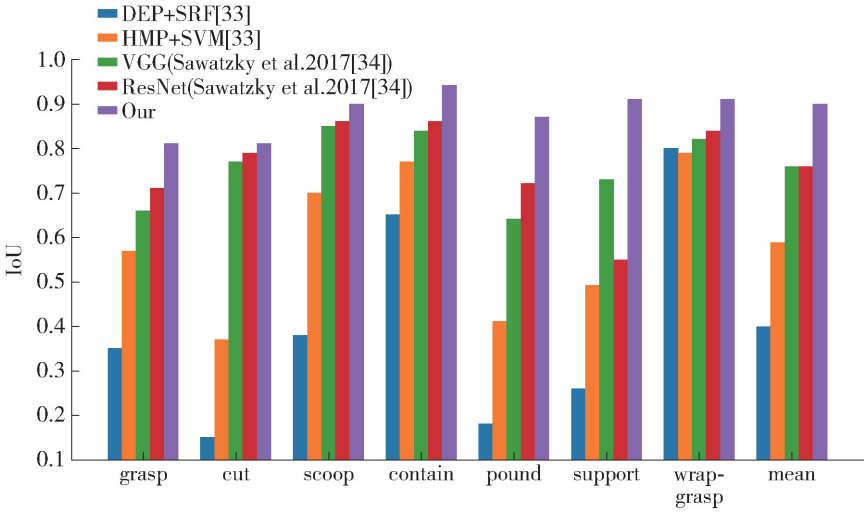


图 7 IoU 度量的 affordance 检测结果

Fig. 7 Performance of the affordance detection with metric of IoU

决策网络处理感知信息.

2.3 记忆网络实现结果

为了帮助机器人理解抓取常识图中的节点和关系,训练了一个包含 1 个嵌入层和 2 个 r-GCN 层的网络. 网络的输入是自建的抓取常识图,其中事实以三元组的形式表示,例如(pour,need,contain)和(scissor,has,cut). 在嵌入层中,一个词向量的维度被设置为 100. 使用 Adam 优化器,将其学习率设置为 0.01,并将每一层 r-GCN 的 dropout 率设置为 0.1. 同时使用了惩罚参数设置为 0.02 的 L2 正则化. 对于每个测试三元组,其头部实体被删除,然后轮流由字典中每个实体替换同时计算得分,并将得分按照降序排列,得分最高的实体被选择作为最终的记忆输出. 记忆网络最终的平均倒数排名(mean reciprocal rank,MRR)为 0.77,并且 hits@10 训练后能达到 0.97. 结果表明记忆网络可以从向量的角度实现对节点和关系的语义理解,并可以根据记忆线索对相关节点或关系进行关联.

2.4 决策网络实现结果

在训练中,决策网络的输入是关于任务、记忆、观测 affordance 和标签的词语,并使用 100 维的嵌入层来处理这些输入. 决策数据集被随机分割成训练集(80%)和测试集(20%). 该决策网络的测试准确率为 99.99%. 测试结果表明,该网络成功地区分了不同的任务,并能够理解对象的 affordance. 使用 6 项常见任务测试了 5 种不同的物体,并将决策结果在总结在了表 3 中. 决策结果的表示形式为:A/B,其中 A 表示任务所需要的 affordance,B 表示要被抓取的物品 affordance. 值得注意的是如果 B 为[none]则表示该物品不能满足任务需求,因此选择不去抓取该物品. 结果表明,该决策网络能够做出准确的决策,即正确地判断一个物品是否可以被操纵执行输入的任务. 如果物品不具有操作任务所需的 affordance,则选择不去抓取该物品,并给出任务所需 affordance 的建议;否则,输出将被抓取的物品 affordance 来指导抓取动作.

表 3 决策网络结果

Table 3 Results of the decision network

物品	挖	舀	切割	敲击	喝	传递
铲子	[support]/[grasp]	[scoop]/[none]	[cut]/[none]	[pound]/[none]	[contain]/[none]	[wrap-grasp]/[support]
杯子	[support]/[none]	[scoop]/[none]	[cut]/[none]	[pound]/[none]	[contain]/[grasp]	[wrap-grasp]/[wrap-grasp]
刀	[support]/[none]	[scoop]/[none]	[cut]/[grasp]	[pound]/[none]	[contain]/[none]	[wrap-grasp]/[cut]
锤子	[support]/[none]	[scoop]/[none]	[cut]/[none]	[pound]/[grasp]	[contain]/[none]	[wrap-grasp]/[pound]
勺子	[support]/[none]	[scoop]/[grasp]	[cut]/[none]	[pound]/[none]	[contain]/[none]	[wrap-grasp]/[scoop]

2.5　认知模型评估

认知模型将 3 个训练好的网络融合在一起，并使用语义向量的形式传递信息. 为了验证认知模型，在测试集的各类型物品中分别选择了 15 张图片进行测试，总共使用 255 张照片作为素材进行抓取决策推理. 为了保证物品 affordance 的完整性，选择的图片中物品的 affordance 均被完整地展示. 如表 4 所示，模型实现抓取决策的准确率为 99.8%，除了其中的 2 个错误决定：抹刀在挖的任务中和锯在敲击的任务中各出现了一次错误决定，查验各环节结果显示是因为模型在感知部分输出的 affordance 产生了误判，以至于输出错误的 affordance 类型. 为了输出给决策网络使用，在感知网络的后处理部分使用了超参数作为像素阈值，对分割出的 affordance 像素数量进行了约束，以保证网络输出的鲁棒性. 大于该阈值则输出该 affordance 类型，否则不会输出. 上述超参数的设置会过滤掉感知网络中误判的 affordance（误判像素数量小于阈值），提高了输出的准确性，同时也会使得部分像素较少的 affordance 特征被过滤，因此输出了有缺失的 affordance 种类，直接影响了后续的决策部分. 认知模型的决策结果可视化如图 8 所示. 橘色框左边的表示输入的任务示意图，橘色框中的图片分别表示模型根据不同任务

表 4　模型测试准确率

Table 4　Test accuracy of the model　%

物品	挖	舀	切割	敲击	喝	传递
碗	100	100	100	100	100	100
茶杯	100	100	100	100	100	100
锤子	100	100	100	100	100	100
刀	100	100	100	100	100	100
长柄勺	100	100	100	100	100	100
木槌	100	100	100	100	100	100
马克杯	100	100	100	100	100	100
罐子	100	100	100	100	100	100
锯	100	100	100	**93.3**	100	100
剪刀	100	100	100	100	100	100
勺	100	100	100	100	100	100
修剪刀	100	100	100	100	100	100
汤匙	100	100	100	100	100	100
嫩化剂	100	100	100	100	100	100
铁铲	100	100	100	100	100	100
抹刀	**93.3**	100	100	100	100	100
锅铲	100	100	100	100	100	100

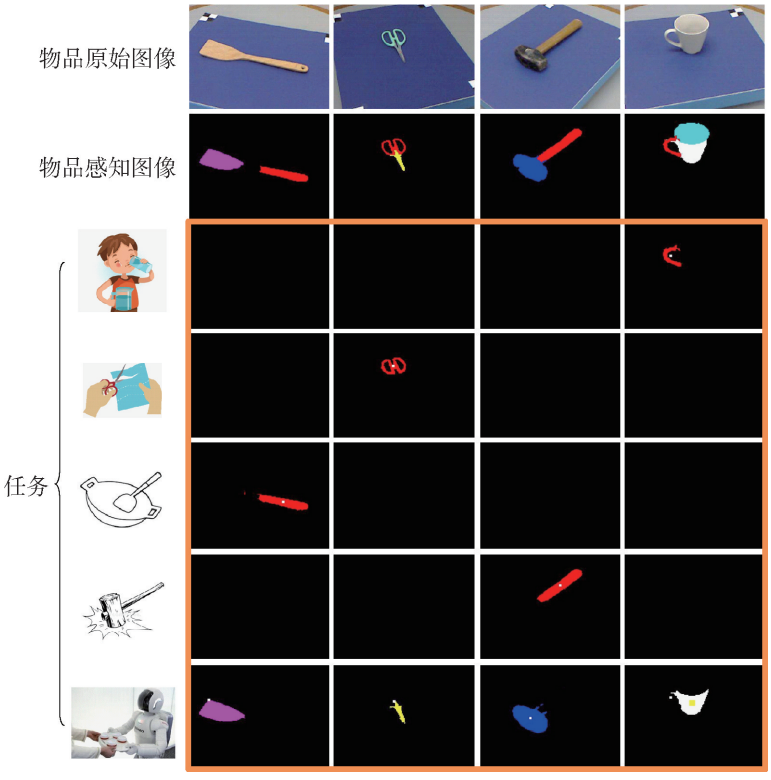


图 8　认知模型的决策结果可视化

Fig. 8　Visualization of the decision results of the cognitive model

得到的抓取位置. 黑色方块代表该物品不适合该任务,因此选择不去抓取. 注意,在可抓取位置中,标记了一个 6×6 的像素块来表示初始抓取位置. 准确率结果证明了认知模型实现了合理灵活的决策. 认知模型以 affordance 的形式实现对物品的基元理解,并通过记忆数据将物品与任务联系起来,从而输出满足任务要求的抓取决策,为后续动作执行提供可靠的初始抓取位置.

3 结论

1) 提出了一个机器人抓取决策的认知模型. 认知决策模型受大脑中分区分块的功能结构的启发,由卷积感知网络(受视觉腹侧信息通路功能启发)、记忆图网络(受海马体信息通路功能启发)和贝叶斯决策网络(受皮层柱信息通路功能启发)三部分组成. 模块化结构使认知模型具有很强的鲁棒性,3 个模块的结构设计和模块之间的协调具有很强的可解释性.

2) 建立了抓取相关的常识图和抓取决策数据集. 在该模型中,将常识图中的物品属性、任务和物品编码为空间向量,以实现语义理解. 对物品、任务、记忆间的关系进行建模,以决策抓取位置.

3) 该模型对 UMD part affordance 数据集的抓取决策准确率达到 99.8%.

参考文献:

[1] MAHLER J, MATL M, SATISH V, et al. Learning ambidextrous robot grasping policies [J/OL]. Science Robotics, 2019, 4(26) [2020-12-01]. http://robotics.sciencemag.org/content/4/26/eaau4984.

[2] ZENG A, SONG S, LEE J, et al. Tossingbot: learning to throw arbitrary objects with residual physics [J]. IEEE Transactions on Robotics, 2020, 36(4): 1307-1319.

[3] ZENG A, SONG S, YU K T, et al. Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching [C] // 2018 IEEE International Conference on Robotics and Automation. Piscataway: IEEE, 2018: 1-8.

[4] LECUN Y, BOSER B, DENKER J S, et al. Backpropagation applied to handwritten zip code recognition[J]. Neural Computation, 1989, 1(4): 541-551.

[5] GIBSON J J. The ecological approach to visual perception: classic edition[M]. London: Psychology Press, 2014: 91-212.

[6] ARDÓN P, PAIRET È, LOHAN K S, et al. Affordances in robotic tasks—a survey [J]. ArXiv Preprint ArXiv, 2020: 2004.07400.

[7] RUIZ E, MAYOL-CUEVAS W. Geometric affordance perception: leveraging deep 3d saliency with the interaction tensor[J]. Frontiers in Neurorobotics, 2020, 14: 45.

[8] OSIURAK F, ROSSETTI Y, BADETS A. What is an affordance? 40 years later [J]. Neuroscience & Biobehavioral Reviews, 2017, 77: 403-417.

[9] DO T T, NGUYEN A, REID I. Affordancenet: an end-to-end deep learning approach for object affordance detection [C] // 2018 IEEE International Conference on Robotics and Automation. Piscataway: IEEE, 2018: 1-5.

[10] NGUYEN A, KANOULAS D, CALDWELL D G, et al. Object-based affordances detection with convolutional neural networks and dense conditional random fields[C] // 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway: IEEE, 2017: 5908-5915.

[11] KOKIC M, STORK J A, HAUSTEIN J A, et al. Affordance detection for task-specific grasping using deep learning [C] // 2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids). Piscataway: IEEE, 2017: 91-98.

[12] CHU F J, XU R, SEGUIN L, et al. Toward affordance detection and ranking on novel objects for real-world robotic manipulation[J]. IEEE Robotics and Automation Letters, 2019, 4(4): 4070-4077.

[13] ZENG A. Learning visual affordances for robotic manipulation [D]. Princeton: Princeton University, 2019.

[14] ARDÓN P, PAIRET È, PETRICK R P A, et al. Learning grasp affordance reasoning through semantic relations [J]. IEEE Robotics and Automation Letters, 2019, 4(4): 4571-4578.

[15] ANTANAS L, MORENO P, NEUMANN M, et al. Semantic and geometric reasoning for robotic grasping: a probabilistic logic approach [J]. Autonomous Robots, 2019, 43(6): 1393-1418.

[16] FANG K, ZHU Y, GARG A, et al. Learning task-oriented grasping for tool manipulation from simulated self-supervision[J]. The International Journal of Robotics Research, 2020, 39(2/3): 202-216.

[17] KARAOGUZ H, JENSFELT P. Object detection approach for robot grasp detection [C] // 2019 International Conference on Robotics and Automation. Piscataway: IEEE, 2019: 4953-4959.

[18] KASAEI S H, SHAFII N, LOPES L S, et al. Interactive open-ended object, affordance and grasp learning for

robotic manipulation[C]//2019 International Conference on Robotics and Automation. Piscataway: IEEE, 2019: 3747-3753.

[19] MISHKIN M, UNGERLEIDER L G, MACKO K A. Object vision and spatial vision: two cortical pathways [J]. Trends in Neurosciences, 1983, 6: 414-417.

[20] AMEDI A, MALACH R, HENDLER T, et al. Visuo-haptic object-related activation in the ventral visual pathway[J]. Nature Neuroscience, 2001, 4(3): 324-330.

[21] KRAVITZ D J, SALEEM K S, BAKER C I, et al. The ventral visual pathway: an expanded neural framework for the processing of object quality[J]. Trends in Cognitive Sciences, 2013, 17(1): 26-49.

[22] ERDOGAN G, CHEN Q, GARCEA F E, et al. Multisensory part-based representations of objects in human lateral occipital cortex[J]. Journal of Cognitive Neuroscience, 2016, 28(6): 869-881.

[23] OSIURAK F, JARRY C, LE GALL D. Grasping the affordances, understanding the reasoning: toward a dialectical theory of human tool use[J]. Psychological Review, 2010, 117(2): 517.

[24] VAN LEEUWEN L, SMITSMAN A, VAN LEEUWEN C. Affordances, perceptual complexity, and the development of tool use[J]. Journal of Experimental Psychology: Human Perception and Performance, 1994, 20(1): 174.

[25] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2015: 3431-3440.

[26] KRISTJÁNSSON Á, CAMPANA G. Where perception meets memory: a review of repetition priming in visual search tasks[J]. Attention, Perception & Psychophysics, 2010, 72(1): 5-18.

[27] EICHENBAUM H, YONELINAS A P, RANGANATH C. The medial temporal lobe and recognition memory[J]. Annu Rev Neurosci, 2007, 30: 123-152.

[28] STARESINA B P, REBER T P, NIEDIEK J, et al. Recollection in the human hippocampal-entorhinal cell circuitry[J]. Nature Communications, 2019, 10(1): 1-11.

[29] MILLER T D, CHONG T T J, DAVIES A M A, et al. Human hippocampal CA3 damage disrupts both recent and remote episodic memories[J]. Elife, 2020, 9: e41836.

[30] SCHLICHTKRULL M, KIPF T N, BLOEM P, et al. Modeling relational data with graph convolutional networks [C]// European Semantic Web Conference. Berlin: Springer, 2018: 593-607.

[31] YANG B, YIH W, HE X, et al. Embedding entities and relations for learning and inference in knowledge bases [J]. ArXiv Preprint ArXiv, 2014: 1412.6575.

[32] TISCHBIREK C H, NODA T, TOHMI M, et al. In vivo functional mapping of a cortical column at single-neuron resolution[J]. Cell Reports, 2019, 27(5): 1319-1326.

[33] ROY A. The theory of localist representation and of a purely abstract cognitive system: the evidence from cortical columns, category cells, and multisensory neurons[J]. Frontiers in Psychology, 2017, 8(187): 186.

[34] SOHN K, LEE H, YAN X. Learning structured output representation using deep conditional generative models [C]// Advances in Neural Information Processing Systems. San Francisco: Margan Kaufmann, 2015: 3483-3491.

[35] KINGMA D P, WELLING M. Auto-encoding variational bayes[J]. ArXiv Preprint ArXiv, 2013: 1312.6114.

[36] MYERS A, TEO C L, FERMÜLLER C, et al. Affordance detection of tool parts from geometric features [C]//2015 IEEE International Conference on Robotics and Automation. Piscataway: IEEE, 2015: 1374-1381.

[37] SAWATZKY J, SRIKANTHA A, GALL J. Weakly supervised affordance detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 2795-2804.

（责任编辑　杨开英）