新型神经网络实现类人概念形成、理解与交流

作者：writer 来源：科学网

本文原地址：https://www.iikx.com/news/progress/40038.html

新型神经网络实现类人概念形成、理解与交流。人类智能的一个独特能力是能够从感官体验中抽象出概念，从而可以脱离感官体验，直接在概念空间进行思考和交流。一般认为，这种将高维感知压缩为低维概念，再由概念重构感知的双向过程构成了人类符号化思维的基础，进而支持了语言的产生。然而，当前的人工智能（AI）系统却难以实现这一过程：传统的深度网络往往将知识纠缠在海量的参数中，难以提取出独立的概念；而受到广泛关注的AI大模型则高度依赖人类已有的语言符号进行训练，无法真正从无到有地从感知经验中自发形成概念。这是当前AI与人脑之间的一个关键差别。

近日，中国科学院自动化研究所脑图谱与类脑智能实验室余山团队与北京大学心理与认知科学学院毕彦超团队的合作研究为解决这一难题提供了重要突破。该研究提出了一种新型神经网络框架CATS Net，实现了类人的概念形成、理解和交流。有趣的是，神经网络自发形成的概念空间与人类语言构成的概念空间有明显的相似性，而且对于这些概念的表征也与人脑内的表征显著相关。这一研究为理解人类的概念认知提供了计算模型，同时为建立具有类人概念智能的人工智能系统打下了基础。相关成果已在线发表于国际学术期刊《自然·计算科学》（Nature Computational Science）。

该研究提出的CATS Net包含两个核心模块：概念抽象（CA）模块与任务求解（TS）模块。在处理视觉任务时，CA模块能够自发地将高维的视觉输入压缩成紧凑的低维概念向量。随后，这些概念向量如同开锁的钥匙一般，通过分层门控机制产生一系列开关信号，可动态调节TS模块的神经网络活动，高效灵活地指导其完成特定的视觉感知任务。系统可以根据与环境的互动自主生成大量新概念，并形成自己的概念空间。当不同神经网络所生成的概念空间对齐之后，就可以不用从环境中学习，而是直接通过概念向量在网络间传递知识。这些能力完整地再现了人脑的概念生成、理解和交流。

进一步，研究团队将　CATS Net　自发形成的概念表征与人类的概念空间和神经活动数据进行了对比。功能磁共振成像（fMRI）的表征相似性分析（RSA）显示，CATS Net　生成的概念空间不仅与心理学上的人类认知语义模型高度一致，其表征模式还与人类大脑中负责视觉语义表征的腹侧枕颞皮层活动模式显著相关。同时，CA模块的动态门控机制则与脑中负责概念提取与操控的语义控制网络活动模式相吻合。这表明，CATS Net不仅在功能层面模拟了人类的概念认知，同时也在机制层面揭示了人脑概念形成与理解的计算原理。

CATS Net来源于前额叶启发的情境化信息处理模型（CDP），这也提示了前额叶和CDP可能在人类概念认知中发挥了核心的作用。该工作为研发具备人类概念形成与应用能力的下一代智能系统奠定了重要基础。当前，大语言模型能力仍受限于人类语言所限定的范畴，赋予他们自主形成

新概念的能力有望促进其在更广阔的领域发挥作用，比如从事全新的科学探索。当然，拥有了这种能力之后，如何确保这些系统与人类的价值对齐，将成为下一步要解决的关键问题。

本研究的共同第一作者为中国科学院自动化研究所博士研究生郭良轩、北京大学博士研究生陈昊扬及中国科学院自动化研究所陈阳副研究员；中国科学院自动化研究所余山研究员、北京大学毕彦超教授与中国科学院自动化研究所陈阳副研究员为共同通讯作者。研究工作得到中国科学院基础研究领域青年团队计划、国家自然科学基金委、中国科学院战略先导专项等资助。（来源：中国科学院自动化研究所）

相关论文信息：https://doi.org/10.1038/s43588-026-00956-4