



观点:

论视觉理解

潘云鹤^{1,2}

¹浙江大学计算机科学与技术学院人工智能研究所, 中国杭州市, 310027

²之江实验室, 中国杭州市

E-mail: panyh@zju.edu.cn

本文译自 Pan, 2022. On visual understanding. *Front Inform Technol Electron Eng*, 23(9):1287-1289.
<https://doi.org/10.1631/FITEE.2130000>

1 视觉识别面对的问题和发展方向

人工智能 (artificial intelligence, AI) 从一开始就将模式识别作为其重要工作领域之一。近年来, 深度神经网络 (deep neural network, DNN) 的成熟使得视觉识别的准确率大为提高, 并在医学图像识别、车辆识别、人脸识别等方面广泛投入应用, 推动 AI 产业发展进入高潮。但是, 基于 DNN 技术的视觉识别存在重要缺陷, 例如, 需要标识的样本数据量较大、模型的鲁棒性较差、学习与推理的结果难以解释、模型泛化与跨数据域迁移能力较弱等。这些问题已成为当前 AI 研究和应用发展的一大挑战, 甚至是障碍。

如果我们用更大的视野来审视当前的视觉识别技术, 可以发现, 上述缺陷是基础性的。原因在于, 当前使用的 DNN 模型在用于视觉识别过程前, 首先需要用大量标识过的视觉数据来训练。从本质上看, 是在进行基于数据统计和模式匹配的分类过程 (Krizhevsky et al., 2017), 所以对样本数据域有很强的依赖。要想具备解释和迁移功能, 仅仅做视觉分类是不够的, 还需经过进一步的视觉理解。

2 视觉理解的三步模型

视觉识别不等同于视觉理解, 它只是理解的第 1 步。视觉理解共有 3 步, 分类是其第 1 步。在

分类之后应进入到第 2 步: 视觉解析。在视觉解析过程中, 应当对视觉对象的各组成部分及彼此结构关系作进一步辨认与比较。辨认是根据已知视觉概念的部件与结构, 寻找视觉数据中对应的各部件与结构的过程。其结果一方面验证分类结果的正确性, 另一方面建立起视觉对象数据的结构。完成视觉解析之后, 还要进入第 3 步: 视觉模拟。在视觉模拟过程中, 要对视觉对象的结构进行包括因果推理在内的预测性运动模拟和操作, 以判断其在现实空间中满足物理约束的合理性, 从而对前面的识别和分析结果加以验证。

现在以一只猫的照片为例, 说明视觉理解的过程模型如下:

1. 识别: 是只猫。取出猫的视觉概念, 进入下一步, 否则结束。

2. 解析: 根据视觉概念所含的结构, 分别辨认猫的头、身、脚、尾及彼此关系是否与猫的概念相适合, 如不适合, 返回第 1 步重新识别; 如适合, 进入下一步。

3. 模拟: 模拟猫的各种活动, 考察猫的活动及其与环境的交互是否能以合理方式完成。如不能, 返回第 2 步; 如能, 进入下一步。

4. 结束理解, 将处理后的结构化数据纳入猫的知识。



潘云鹤教授

FITEE 主编

3 视觉理解模型的特点

为进一步认识上述视觉理解的3步过程模型，我们对其若干特点作进一步阐述：

1. 视觉理解的关键步骤是视觉解析。视觉解析是一个根据视觉识别得到的视觉概念（Pan, 2019），按照概念结构所含各部件，对理解对象作相应辨认。解析一个视觉理解对象，按照从顶向下的次序，是一个从概念树的根开始，到枝，直至到叶的对视觉数据的辨认与建构的过程。

2. 人类的视觉解析任务往往只针对构成概念的主要部件。主要部件都各有对应的语言名称进行描述。而对于尚未用语言描述的非主要部件，如面孔的颧骨与下巴间的部位，只有受过专门解剖训练的专家（如医生、画家等）才具有专业概念与记忆。所以视觉解析会同时处理视觉与语言等多重知识（Yang et al., 2008），是一个跨媒体（Pan, 2020b）处理的过程。

3. 视觉知识（Pan, 2019）对视觉解析和视觉模拟都是必不可少的，因为它提供的部件结构为部件的辨认与比较提供了明确指向。每个人在视觉知识建立之初，都要感谢父母与老师。当他们对孩子讲：“看，这是小猫。小猫有尖耳朵、圆眼睛、长胡须、四条短腿跑得很快、跳得很远，一下抓到老鼠。”他们就在引导孩子学会如何在长期记忆中建构起基本的视觉知识。

4. 经过理解的视觉数据，实际上已经被结构化，从而形成视觉知识。这样的视觉知识也很容易被纳入长期记忆。例如你看到一只猫的头特别小，或者其毛色与斑纹特殊，或其行走姿态特殊，都可能以扩大“猫”的概念范畴（Pan, 2019）而被纳入你的记忆中的“猫”的视觉概念。视觉概念的范畴很重要，其疆域大小反映了见多识广的程度。实际上，用于学习DNN模型的样本数据并非越多越好，而是在概念范畴的对应中分布得越广泛、越均衡才越好，因为基于这样的样本数据所训练出的模型的鲁棒性和泛化能力才越强。

5. 理解后的视觉信息当然就可以被解释了，因为它已具备深层结构认知；也可以用于迁移学习，因为它已具备跨媒体的语义概念。这些语义信息能清晰地指出识别迁移的合理方向。

4 将视觉识别推进到视觉理解

视觉理解是重要的，因为它与视觉知识（Pan, 2019）、多重知识表达（Pan, 2020b）等共同作用，将会为AI研究打开一扇新大门。视觉理解既是视觉识别的深入，又是视觉知识学习（Pan, 2020a）与运用的深入。AI研究视觉识别的时间之长已超过半个世纪。与之并行起步的语音识别，早已进入分析单词、句子和段落的语言理解，由此成功发展了人机对话和机器翻译，取得众所周知的里程碑式重大进展。因此，我们建议将视觉识别推进到视觉理解，现在正是瞄准这一更深刻的视觉智能行为的最佳时机。

致谢

感谢浙江大学庄越挺、吴飞、耿卫东、杨易、孙凌云、汤斯亮等教授为本文提供富有价值的建议。

参考文献

- Krizhevsky A, Sutskever I, Hinton GE, 2017. ImageNet classification with deep convolutional neural networks. *Commun ACM*, 60(6):84-90. <https://doi.org/10.1145/3065386>
- 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*, early access. <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>
- Yang Y, Zhuang YT, Wu F, et al., 2008. Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval. *IEEE Trans Multim*, 10(3):437-446. <https://doi.org/10.1109/TMM.2008.91735>