中国科学院大学课程讲义

课程： 计算机视觉

（第三章）

教 研 室： 脑科学与智能技术

教 师： 胡占义

编写时间： 2018-9-14

讲义几点说明

本章为国科大硕士研究生春季学期开设的《计算机视觉》课程讲义的第三章，旨在对灵长类动物的神经物体表达（neural object representation）进行简单介绍。物体表达是物体识别的基础，也是视觉感知（visual perception）和视觉认知（visual cognition）之间的桥梁。深度学习在图像物体分类的成功，也很大程度上在于深度网络学习到了一种“抽象的物体表达”，如 AlexNet, VGG 的卷积层和全连接层（非分类层）的输出，可以视作对图像特征和物体的一种表达。所以，计算机视觉研究人员了解一些生物物体表达内容是有益和必要的。《计算机视觉》为 40 学时的专业普及课，授课老师为： 胡占义，董秋雷，申抒含。本章为 胡占义一人撰写。

随着互联网的普及，人们对教材与参考文献阅读的习惯已发生了本质的变化。现在似乎已很少有人再仔细研读一本教材，而大家往往是根据需要，从网上寻找“合适的具体内容”。所以，为了大家阅读参考方便，《计算机视觉》的课程讲义也以单章形式给出。

目前几乎任何一所高校都有从事计算机视觉的研究人员，但很多学生，包括老师，大都没有系统上过计算机视觉课，特别是“深度学习热潮”前的相关内容。笔者觉得，目前很多计算机视觉研究人员似乎连计算机视觉的奠基者：David Marr, 及其提出的计算视觉理论也很少有人知道了。为了给相关人员提供一些参考和帮助，同时也作为一名科研人员回报社会的方式，本教程讲义完成后，已放在笔者课题组主页上供大家免费下载阅读。

http://vision.ia.ac.cn/zh/progress.html。

该讲义为笔者 30 多年来从事计算机视觉研究的一些心得和总结，不妥之处请大家批评指正。笔者长期以来得到国家自然科学基金委、科技部、中科院和国科大的资助，在此一并表示感谢。

2018-11-20

中国科学院自动化研究所/模式识别国家重点实验室

第三章：猴子与人视觉皮层中的神经物体表达及其基于 DCNN 的建模

摘要

人在日常生活中可以毫不费力地对物体进行识别，并且对视角和光照变化具有很好的适应性。然而，视角和光照变化却是阻碍计算机视觉中物体识别，乃至整个计算视觉理论的主

要困难。所以，探索生物（特别是灵长类动物）的物体识别的机理，不仅对于揭示生物视觉奥秘具有重要意义，而且对于启迪和推动视觉计算方法和理论具有重要意义。

物体表达是物体识别的基础和核心科学问题。视皮层中的物体表达，就是指视觉皮层中群体神经元在图像物体刺激下的“响应（发放脉冲）模式”。物体表达是“视觉物体感知”和“视觉物体认知”之间的桥梁。

人类这种看上去毫不费力的物体识别能力，其神经加工机理至今却仍是神经科学领域的一项尚未有明确结论的重要研究内容。猴子和人对图像物体的表达机理是否相同？如何进行相互之间的比较对比？视皮层中图像物体的神经表达方式是什么？深度卷积神经网络（DCNN）是否可以作为视皮层中的物体神经表达的框架性建模方法？本章将对文献中这方面近年来的一些相关进展进行简单介绍，以期对相关人员具有一些帮助和参考价值。

本章内容是中国科学院大学“计算机视觉”研究生课程讲义的一部分。为了便于“非神经科学领域”的读者了解相关内容，本章将重点介绍“主要思想和结论”，而不是对具体方法和细节进行介绍。另外，由于笔者不是研究神经科学的，所以不论对内容的理解还是总结介绍均不可避免地有不严格甚至错误的地方，在此一并表示歉意。

3.1 生物视觉物体识别中的“基本物体识别”问题

生物对图像物体的识别，可以分为“基本物体识别”（core object recognition）[1]和一般物体识别。所谓的基本物体识别是指在短时间内（物体出现后100毫秒左右）即可完成的识别。神经科学界普遍认为，在这样短的时间内，视觉皮层的高层区域到底层区域的反馈（feedback）效果不会很大，因此，视觉基本物体识别的神经加工机制大体可以认为是一个“前馈（feed forward）”加工过程。一般物体识别涉及注视、工作记忆、语义融合等内容，一般涉及皮层不同区域之间的反馈和抑制，而这些内容即使在神经科学领域，目前也缺乏足够的机理性描述，所以，本章后面关于基于DCNN对物体的表达建模部分，仅仅介绍“基本物体识别”下的物体表达建模问题。

这里强调一下“recurrent”neural network 这个概念，信息领域对“recurrent”这个词的使用似乎有点模糊。 “recurrent” 包含两方面的内容：一是指网络含有同层之间的抑制(lateral inhibition), 另一种是指含有高层对底层的反馈（feedback）。不少文献中给出的神经网络仅仅有同层之间的抑制，也统称为“Recurrent NN”，可能描述得更具体一点会有助于读

[1] Core visual object recognition： Visual object recognition within one fixation without contextual influence, eye movements, or shifts in attention

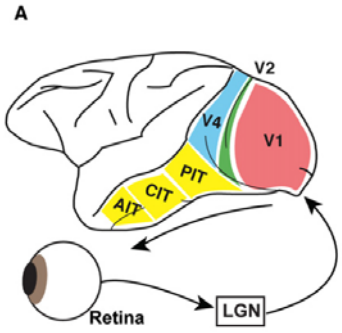
者对内容的准确理解。

3.2 物体识别通道 –腹部通道（visual ventral pathway）

2012 年 DiCarlo 等（DiCarlo et al. 2012）在 Neuron 给出了基本物体识别的“解缠（untangling）”理论。主要思想为：视觉腹部通道的功能，在于去除物体在视网膜上成像过程中导致的“纠缠现象”。所谓“纠缠现象”，就是指不同物体的表达无法“线性可分”。如在图像空间的二个不同物体很难用线性分类器直接进行分类。

DiCarlo 等认为，图 3.1(a) 中的腹部通道，从视网膜（retina）→V1→V2→V4→PIT→CIT→AIT 的逐层加工过程，旨在去除“纠缠现象”。到 IT 区（Inferior temporal cortex），“图像的纠缠”已去除，IT 神经元对物体的表达已线性可分。

图 3.1(b)中，Dicarlo 等给出了猴子视皮层不同区域的输入、输出和内部神经元的个数。从图 1(b)可见，同一皮层区域内部神经元的个数远远多于该区域输入和输出神经元的个数，说明同一皮层区域内部神经元的相互作用是信息加工的主体。另外，输入输出神经元的数目也比当前的 DCNN 多很多。这些都是猴子腹部通道（物体识别通道）与当前 DCNN 的一些不同之处。另外，当前 DCNN 的卷积层中，不同神经元的相互作用仅仅为简单的“卷积操作”，缺乏同层抑制。



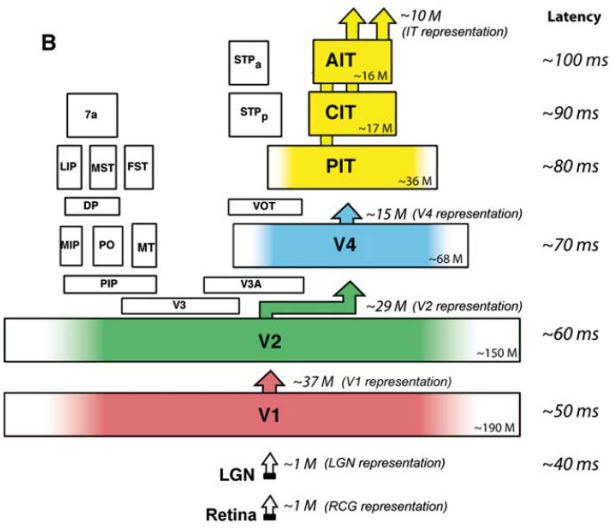


图 3.1(A)： 猴子的视觉腹部通道; 图 3.1(B): 不同皮层区域的大小（方框大小）、神经元个数（输入、输出、内部神经元）和潜伏期（latency）(摘自 DiCalo et al. 2012)

3.3 视觉皮层中的物体表达

视觉皮层中的物体表达，是指某个皮层区域中群体神经元对某个刺激物体的脉冲发放模式。如 V4 区神经元对物体的表达，是指 V4 区群体神经元对该物体刺激下的脉冲发放模式。单个神经元的表达能力很弱（也不排除有些特殊神经元对特定物体放电），一般来说，群体神经元对某个物体刺激下的放电模式称之为该物体的神经表达。图 3.2 为某个神经元在光栅刺激的脉冲发放示意图：

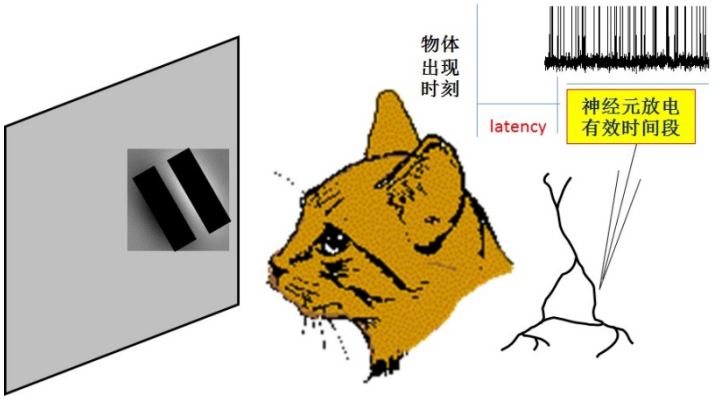


图 3.2：猫 V1 区某个神经元在光栅刺激下的脉冲发放情况。由于脉冲连续发放，表明该神经元对当前光栅方向很敏感

目前的研究表明，利用一段时间内的脉冲发放的个数来表示神经元的响应（response）（称

之为速率（rate）模型），是一种比较有效的神经响应表达（neuron response representation）. 人们并没有发现直接利用脉冲序列来表示神经响应有更多好处。所以，文献中关于神经元的响应模型，一般指这种速率模型。

物体识别的基础是物体表达。物体表达不论在神经科学领域还是计算机科学领域，均是一个重要的科学问题，并具有重要的应用价值。目前的研究表明，猴子和人的 IT 区是物体表达的主体皮层区。

3.4 猴子 IT 区的物体神经表达可以线性分类

如何度量猴子 IT 区的物体表达能力呢？2015 年 Majaj 等（Majaj et al. 2015）采用了如下方法进行度量。首先，假定：**猴子与人具有相同的物体表达机理。即人难识别的物体，对猴子也难。** 在这种假设下，他们利用图 3.3 所示的 8 类物体（每类 8 个不同个体，图像变形从低到高）来比较猴子和人对这些物体的分类和识别精度，进而确定猴子 IT 区神经元的物体表达是否可以线性分类和物体鉴别， 比较原理如图 3.4 所示。

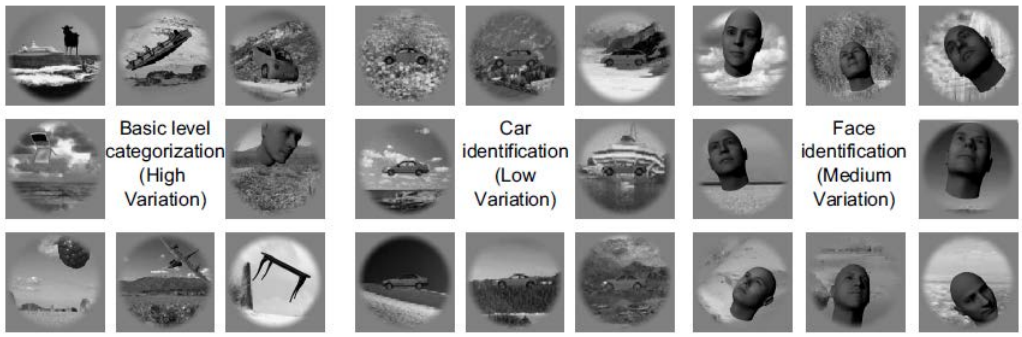


图 3.3：使用的 8 类待分类和鉴别的物体。每类 8 个个体，分类和鉴别难度从易到难。左图为变形下的 8 类物体。中间为低变形对应的 Car 类的 8 个个体；右边为中变形下 Face 类的 8 个个体（摘自 Majaj et al. 2015）

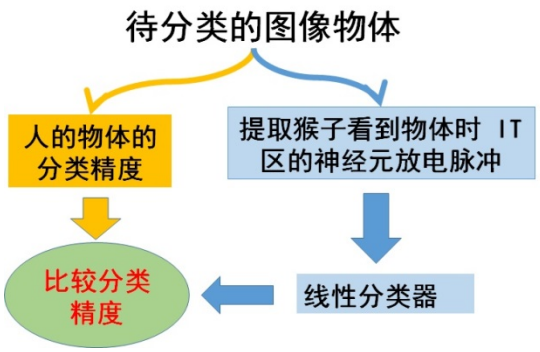


图 3.4：对猴子 IT 区的物体表达是否可以线性可分的测定原理和流程示意图

Majaj 等（Majaj et al. 2015）通过记录二只猴子的 128 个神经位置（neural sites），发现不论是图像分类问题，还是同一类内的个体鉴别问题，对猴子 IT 区的神经表达进行线性分类的结果，都非常接近人类的对应分类结果。由此表明，猴子 IT 区的物体表达，已完成了 DiCarlo 等（DiCarlo et al. 2012）提出的“图像解缠”效应。同时也表明，IT 区确实是图像物体表达的皮层区域。

3.5 人和猴子 IT 区对图像物体的表达具有相似性

上节的假设是：“人难识别的物体，对猴子也难”。那么，猴子与人对物体的表达方式具有相似性吗？2008 年，Kriegeskorte 等（Kriegeskorte et al. 2008）采用图 3.5 所示的途径对猴子和人的物体表达方式进行了相似性比较。如图 5 所示，他们记录了二只猴子 IT 区的 674 个神经元对 92 个不同的图像物体的响应和 4 个人观看同样 92 个物体时的 FMRI 数据（IT 区部分的数据），首先分别计算各自的表达不相似矩阵（RDM：Representational Dissimilarity Matrix）（Nili et al. 2014），然后度量猴子和人的 RDM 之间的相关性。如果相关性强，则表明人与猴子对物体的表达具有相似性，否则表示不存在相似性。

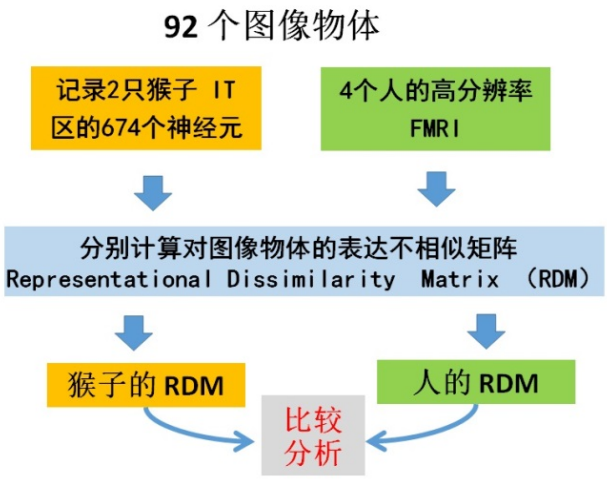


图 3.5：猴子 IT 区与人 IT 区关于物体表达相似性的度量比较方法

RDM 是一个 92*92 的对称矩阵，第(i,j)元素为: $RDM(i,j)=1-r_{i,j}$， 其中$r_{i,j}$表示 674 个神经元（或 4 个人）对第 i 幅图像和第 j 幅图像响应（FMRI）之间的相关系数。图 3.6 为猴子和人对应的 RDM：

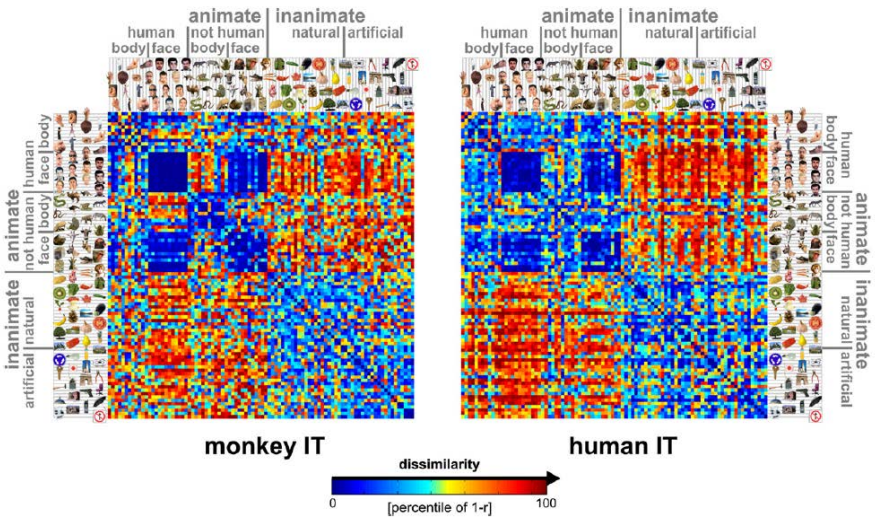


图 3.6： 左图为猴子对应的 RDM，右图为人对应的 RDM。从这两个 RDM 可以看到，它们之间具有很强的相似性（摘自 Kriegeskorte et al. 2008）

Kriegeskorte 等（Kriegeskorte et al. 2008）的结果表明，猴子的 RDM 与人的 RDM 具有很强的相关性，表明猴子的 IT 区的物体表达与人的 IT 区的物体表达具有很强的相似性。

另外，他们利用 MDS （Multidimensional Scaling）算法对猴子的 IT 区神经元响应和人的 FMRI 数据进行了降维分析，结果如图 7 所示。图 3.7 表明，物体的相似度越高，不论对猴子的响应数据还是人的 FMRI 数据，图像越靠近。图 3.7 再次表明人与猴子的图像物体表达具有很高的相似性。

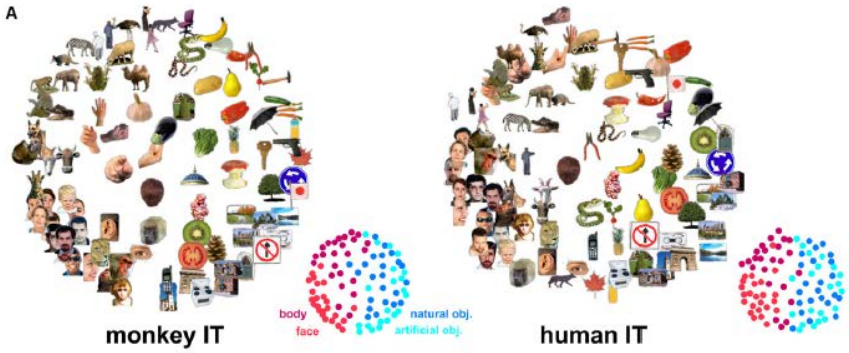


图 3.7：将猴子 IT 区的响应与人的 IT 区的 FMRI 数据利用 MDS 降维到二维空间，发现越相似的物体靠的越近，且猴子数据和人的数据具有相似的排列。结果再次表明猴子和人在 IT 区对图像物体的表达具有相似性（摘自 Kriegeskorte et al. 2008）

另外，他们利用聚类树方法对两种数据分别进行聚类，发现聚类结果，如图 3.8 所示，也具有非常高的一致性。

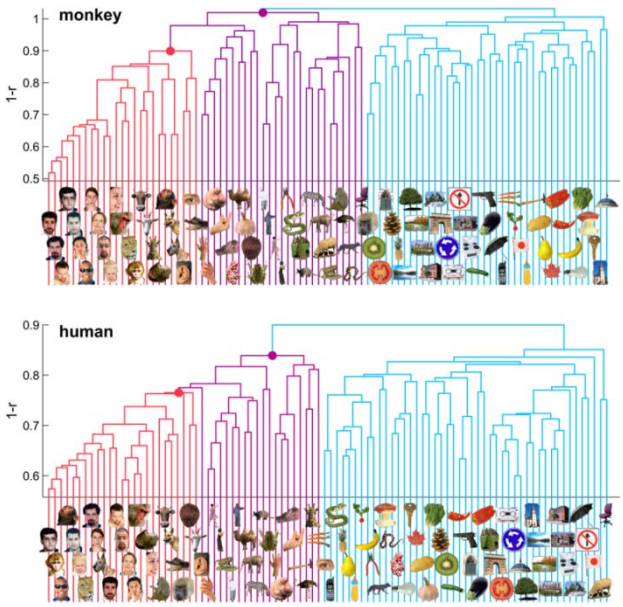


图 3.8：两种数据的聚类结果，表明猴子的物体表达和人类具有很高的相似性 （摘自 Kriegeskorte et al. 2008）

通过上述三种不同方法的比较，Kriegeskorte 等（Kriegeskorte et al. 2008）认为，猴子与人在 IT 区对图像物体的神经表达具有一致性。

3.6 DCNN 在物体表达方面可与猴子 IT 区相媲美

深度网络和深度学习在图像分类和识别方面取得了革命性进展，分类和识别性能已可与人类相媲美，甚至在特定领域超越了人类。那么，深度网络下的物体表达与猴子 IT 区的表达在物体识别方面的性能如何呢？ 2014 年，Cadieu 等（Cadieu et al. 2014）对这个问题进行了分析比较，他们发现从物体表达的观点看，目前深度学习下的物体表达，已与猴子 IT 区的物体神经表达相媲美。

那么如何比较 DCNN 的物体表达与猴子 IT 区的物体神经表达呢？Cadieu 等提出了下面的比较原理和准则：

比较原理：物体表达可以看作是对输入图像的特征提取过程， 如果特征提取的好，则对应预测函数的复杂度(complexity)低.

比较准则: 在同样误差下，给定两种不同的表达，对应预测函数复杂度低的表达好。

Cadieu 等将下面线性回归的正则项系数 λ 的倒数 $\frac{1}{\lambda}$ 作为待分类问题的复杂度，在不同复杂度的物体分类下，比较了多种物体表达方式。发现 Zeller&Fergus (2013) 的深度网络的分类性能几乎与猴子 IT 区表达下的分类性能相媲美. 具体情况如下：

令样本集为 $\{(x_1 \quad y_1),(x_2 \quad y_2),\cdots,(x_n \quad y_n)\}$，其中 x_i为图像物体刺激，y_i为物体的类别标签（如椅子，人等），令函数 $\phi(x)$ 为物体 x 的某种表达，给定某种核函数$\mathrm{k}_\sigma(*,*)$，定义如下的数据矩阵：

$$K_\sigma = \begin{pmatrix} k_\sigma(\phi(x_1),\phi(x_1)) & k_\sigma(\phi(x_1),\phi(x_2)) & \cdots & k_\sigma(\phi(x_1),\phi(x_n)) \\ k_\sigma(\phi(x_2),\phi(x_1)) & k_\sigma(\phi(x_2),\phi(x_2)) & \cdots & k_\sigma(\phi(x_2),\phi(x_n)) \\ \vdots & \vdots & \vdots & \vdots \\ k_\sigma(\phi(x_n),\phi(x_1)) & k_\sigma(\phi(x_n),\phi(x_2)) & \cdots & k_\sigma(\phi(x_n),\phi(x_n)) \end{pmatrix}$$

将正则化的线性回归问题可表示为：

$$\min_{\Theta\in R^n} \frac{1}{2}\|Y - K_\sigma\Theta\|_2^2 + \frac{\lambda}{2}\Theta^T K_\sigma \Theta$$

其中 Θ 为回归模型参数。Cadieu 等(Cadieu F. C. et al., 2014)发现，参数 $\frac{1}{\lambda}$ 确实可以表示回归问题的复杂度。图 3.9 显示，随着 $\frac{1}{\lambda}$ 的增大，分类问题变得越来越复杂，只有深度网络才可以取得比较好的分类结果：

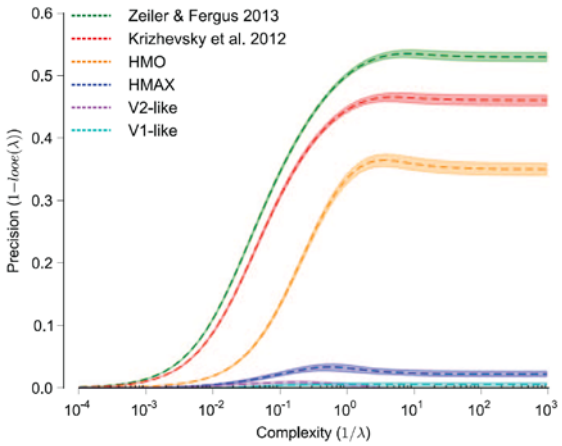


图 3.9：正则项系数 $\frac{1}{\lambda}$ 可以刻画线性回归问题的复杂度(摘自 Cadieu et al., 2014)

图 3.10 表明，当回归问题复杂时，Zeiler & Fergus (2013) 的深度网络可以取得与 IT 神经元表达相似的物体分类性能：

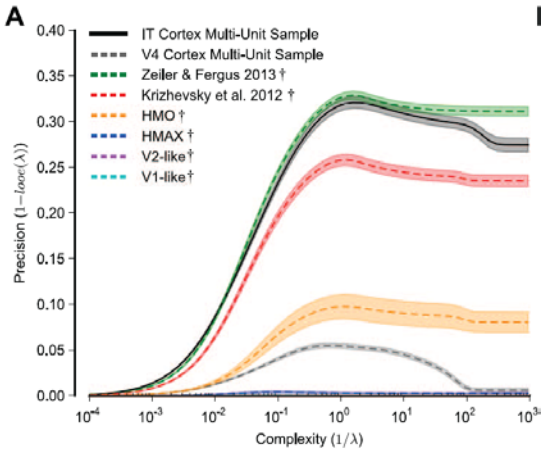


图 3. 10：Zeiler 等的深度网可以与猴子 IT 神经元的物体表达能力相媲美(摘自 Cadieu et al., 2014)

3.7 利用 DCNN 对猴子 IT 区物体表达的建模

（1）物体表达建模的固有困难

研究图像物体的表达问题的困难，在于图像物体很难用特定的参数（维度）来描述。如运动可以用“运动方向和大小”来参数化，但物体很难用有限的维度来刻画，所以物体表达问题是一个困难的问题 （Kourtzi & Connor 2011）。

2017 年，Chang 等(Chang & Tsao. 2017) 表明，一幅正面人脸图像可以用 50 维参数（25 维几何，25 维纹理）来生成。在这种参数化下，他们发现猴子 IT 区的人脸表达，是一种称之为 axis-model 的线性表达方式。即使对人脸这种非常特殊的物体类别（样本差距不大，平面类，少纹理），笔者觉得，axis-model 的有效性也严重依赖于图像必须是“给定 50 维参数分布下” 样本生成的人脸图像（由 200 幅真实人脸图像统计得到的分布）。 很多人脸图像很难用 Chang 等给出的参数分布描述，而对这些不满足该分布的人脸图像，尽管猴子对这些人脸仍在 IT 区进行表达， 但 Axis-model 很难给出很好的预测。

另外一个问题是所谓的“类别表达”和“特征表达”之争。即灵长类的腹部通道到底是表达“类别性息”还是“特征信息”？目前仍没有定论。这个问题，后面将进一步进行介绍。

（2） DCNN 作为框架性建模方法存在的问题

鉴于 DCNN 在图像分类领域取得的巨大成功，以及 DCNN 的层次化结构与视觉腹部通道的层次化信息加工方式具有某种相似性，人们自然会想到：DCNN 可以作为视觉建模的框架性技术（framework）吗？“框架性技术”，这里指一种比较通用的技术，而不是一种针对特

定问题的建模方法。

2015 年，Kriegeskorte（Kriegeskorte N. 2015）在 Annual Review of Vision Science 上撰文称，DCNN 可以作为一种框架性技术来建模生物视觉和脑信息加工过程。2016 年 Kheradpisheh 等（Kheradpisheh et al. 2016）认为，DCNN 与人类不变物体的前向处理过程相似 （Resemble Human Feed-forward Vision in Invariant Object Recognition ）。2016，Yamins & DiCarlo (Yamins & DiCarlo 2016) 在 Nature Neuroscience 撰文提出了“目标驱动的感知建模方法” （The Goal-driven deep learning models to understand sensory cortex）。这些工作均认为， DCNN 可以作为视觉感知的框架性建模方法。

笔者认为，“DCNN 能否作为一种框架性建模方法”，依赖于多个因素。首先，如何评价“建模的性能”？正如近期 Dicarlo 团队（Rajalingham et al. 2018）的工作表明，如果仅仅以物体的“识别率”来度量，则 DCNN 的建模性能已与人和猴子 IT 区的物体表达相近。但如果评价标准更加细化，当既要求 DCNN“对同一类物体不同个体的正确识别率”与人相近，同时又要求**“**类内不同个体误分为其它类**”**的概率也相近，则 DCNN 与人和猴子 IT 的表达还有很大的不同。这也是完全可以理解的。如果任何评价度量下，DCNN 都可以很好建模人和猴子的 IT 区，则意味着人的腹部通道就是一个 DCNN 网络，这将是不可想象的，甚至是荒唐的。

DCNN 能否很好对视觉腹部通道建模的另一个关键因素是 DCNN 的结构问题。DCNN 是一个比较泛的概念，可以有不同的结构（architectures）。如对猴子 IT 区的物体神经表达建模，DCNN 的网络结构应该如何选取？DCNN 需要多少层？需要不需要跨层连接？滤波器的大小如何取？通道的个数应该是多少？等等。我们知道，DCNN 的超参数（hyperparameters）选取问题，是一个比 DCNN 的训练更困难的问题。笔者觉得，选择合适的网络结构，似乎与建模本身的难度没有多少区别。作为一种视觉建模的框架性方法，如果这些关键因素“需要靠实验和技巧”来确定，就很难说 DCNN 具有“框架性技术”应具备的通用性。

另外一个问题是由于采集数据的困难，人和猴的数据量一般不大，而 DCNN 的表达空间很大。这样，利用一个非常大的表达空间来拟合一个小数据，其结论的扩展和推广性值得斟酌。对人和猴子，“大数据”也仅仅是对几百张图像，最多上千张图像刺激下的记录数据。DCNN 对这样数据规模的拟合和预测结果，当网络结构和参数保持不变时，是否对其它数据仍有效，是一个有待进一步深入分析的问题。

DCNN“表达空间”具有的容量与待表达问题之间的差距，会导致一些不同的结论，甚至是一些颠覆性结论。如 2018 年 DeepMind 在 ICLR2018 上的文章指出（Morcos et al. 2018.

ICLR2018 Best paper），选择性高的神经元并不是最重要的神经元。该结论与传统神经科学领域的主流观点和操作相悖。因为在神经科学领域，经典的操作程序为：如要研究 IT 神经元对人脸的表达（编码），首先要确定对人脸具有高选择性（较其它对比实验（control））的 IT 神经元，然后再对这些具有高选择性的神经元开展进一步分析。笔者觉得，Mocros 等（Mocros et al. 2018）得到的结论，可能很大程度上是由于他们使用的 DCNN 的固有容量不大而待分类的 1000 类物体需要大的表达空间所致。由于固有表达空间不大而 1000 类物体需要足够大的表达空间，这样，DCNN 的每个神经元必须对多个物体产生响应，从何降低了 DCNN 神经元对某一特定类的选择性。事实上，仔细分析 Mocros 等结果随 DCNN 层数增大的变化，也可说明这一点。

（3） 基于 DCNN 的视觉建模

文献中倡议 DCNN 可以作为视觉建模的一种框架性方法（Kriegeskorte 2016；Kheradpisheh et al. 2016；Yamins&DiCarlo 2016; Hong et al. 2016; Kietzmannet al. 2017），事实上核心文献是 Yamins 等（Yamins & DiCarlo 2014）在美国科学院院刊上发表的文章：Performance-optimized hierarchical models predict neural responses in higher visual cortex。在这篇文章中，作者发现，对一个 4 层 DCNN （称之为 HMO （Hierarchical Modular Optimization），如图 3.11 所示）仅仅在控制网络分类性能下进行参数学习，学习好的 DCNN 的输出不仅可以定量预测猴子 IT 神经元的脉冲响应，而且 DCNN 倒数第一层的输出同时可以很好预测猴子 V4 区的神经响应。Yamins 等（Yamins & DiCarlo 2014）的工作多少有些令人吃惊。因为在训练该 4 层网络时，控制的仅仅是网络的图像分类性能，并没有利用任何猴子的 IT 神经元的对应输出进行训练，结果不仅最后一层的输出可以很好预测猴子 IT 区神经元的放电，而且倒数第一层的输出同时可以很好预测猴子 V4 区神经元的放电。

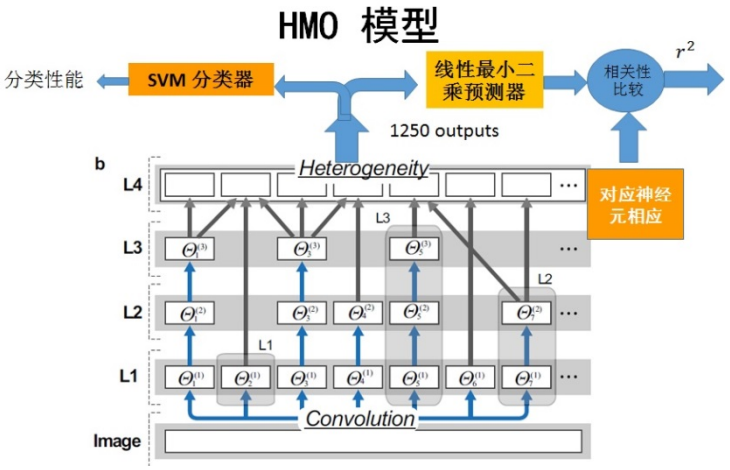


图 3.11： 4 层 DCNN 的最后一层输出可以很好预测猴子 IT 区神经元的放电响应。倒数第一层的输出可以很好预测猴子 V4 区的响应

2016 年，DiCarlo 课题组（ Hong et al. 2016）撰文指出，猴子 IT 区的神经元的响应，同时包含与物体类别无关的信息（identity-orthogonal properties）。通过对一个 6 层的 DCNN 在 ImageNet 数据下仅仅控制图像分类性能进行训练，则训练好的 6 层 DCNN 网络不仅可以很好预测 IT 区神经元对图像物体刺激的响应，而且随着 DCNN 层数的增高，对应的神经元（units）的输出对与类别无关的信息的预测能力也在逐层提高，如对物体的姿态，位置和大小等信息的预测能力。该项工作的意义在于利用计算机视觉界广泛使用的 ImageNet 数据对 6 层网络进行了训练，而训练好的网络则可以定量预测物体的类别信息和与类别无关的其它信息。

另一项也经常用来支撑 DCNN 可以作为视觉建模的一种框架性方法的文献是 2014 年 Kriegeskorte 组（Khaligh-Razavi & Kriegeskorte 2014）在 PLOS Biology 上发表的文章：Deep supervised, but not unsupervised, models may explain IT cortex representation。 该文比较了 27 个简单模型和一个 8 层 DCNN， 发现假如不强制类别信息（如脸谱，动物，体型等 10 类类别信息），DCNN 无法达到猴子 IT 区的表达性能（即 DCNN 的 RDM 与猴子 IT 区的 RDM 之间的相关性小于猴子 IT 区一半神经元对另一半神经元的 RDM 之间的相关性）。当同时利用这些信息时（文中称之为有监督的模型，即 DCNN 第 8 层的输出+第 7 层输出经过 SVM 对动物/非动物，脸/非脸，肢体/非肢体三个类别分别进行分类的输出），则可以达到猴子 IT 区的性能。

在这些工作的基础上， Yamins&DiCarlo 2016（a）在 nature Neuroscience 上提出了基于 DCNN 目标驱动的感知建模方法。该方法认为，视觉腹部通道建模，可以在物体分类性能控制下训练一个 DCNN 完成。下面将对该方法进行介绍。

（4）： 目标驱动的视觉腹部通道建模

Yamins & DiCarlo 2016（a）提出的“目标驱动的视觉腹部通道建模”方法，可以简单表述为：

依据 DCNN 的图像分类性能，在一些控制 DCNN 结构的参数集中选择一个分类性能比较好的 DCNN，进一步在大型图像分类集，如 ImageNet，对该 DCNN 进行分类训练，训练好的 DCNN 则能很好预测猴子的视觉腹部通道性能，即该 DCNN 为腹部通道的优良计算模型。

注意，这里的对腹部通道的建模，不仅要求 DCNN 的最后一层的输出可以很好预测猴子 IT 区的放电响应，前面层的输出应该同样可以预测腹部通道的前面区域的响应，如 V4 以及 V2 区的响应。

Yamins & DiCarlo 2016（a）提出的基于 DCNN 目标驱动的感知建模方法，具有理论上的优美性。不仅适用于视觉建模，同样适用于听觉、触觉等感知通道的建模。

但笔者认为，仅仅靠控制“图像分类性能”训练 DCNN 来对腹部通道进行建模，不管理论多么优美，其普适性仍然需要进一步验证。首先对 DCNN 进行图像分类训练，并没有使用任何关于猴子或人在图像物体刺激下 IT 区的响应信息，如果目标驱动的方法是有效的，那意味着“猴子”和“人”关于物体表达的方式完全相同，这显然与文献中的结果不太符合。因为人 IT 对物体的表达，尽管与猴子 IT 区的表达具有相似性，但仍有很大的不同。另外，正像董秋雷等（Dong et al. 2018）对“目标驱动的视觉腹部通道建模方法”的评论文章所指出的那样，对同一 DCNN 网络结构，在不同的初始化方式下进行训练，对应的 DCNN 的分类结果相近，但表达差异很大，说明目标驱动的视觉建模方法缺乏唯一性。缺乏唯一性说明方法本身存在一些固有缺陷。事实上，Dicarlo 自己课题组 2018 年的文章（Rajalingham et al. 2018），也说明当评价尺度细化后，目标驱动下基于 DCNN 的建模方法，仍无法对猴子的 IT 区很好建模。笔者甚至觉得，文献中经常引用的 Kriegeskorte 组（Khaligh-Razavi & Kriegeskorte 2014）用来支持 DCNN 可以作为视觉建模的框架性方法的文献，本质上也说明目标驱动的建模方法有缺陷。因为不使用明确的 SVM 分类器得到的分类性息， DCNN 分类下的 RDM 与猴子 IT 区的 RDM 有显著差距，说明目标驱动的建模方法，即使用比较粗糙的物体分类下的 RDM 之间 的相关性系数进行度量，仍存在显著差距。

3.8、物体神经表达中的类别信息，形状信息和物体的感知相似性

从上节可知，物体类别性息在物体神经表达中起重要的作用，对目标驱动的视觉建模（Yamins&DiCarlo 2016（a））而言，起到了决定性作用。文献中报道，对基于 DCNN 的场景识别中（Cichy et al. 2015），即使没有使用任何物体类别信息，DCNN 神经元仍对类别具有一定的选择性。更令人吃惊的是，即使 DCNN 的权重随机赋值，对应的 DCNN 仍对物体类别具有一定的选择性。这说明 DCNN 的层次结构会自然导致物体类别信息的形成。但笔者课题组重复这个实验时，发现对 DCNN 的权重随机赋值，似乎 DCNN 的输出并没有包含显著的“类别信息”。

Bracci 等（Bracci et al. 2016）设计了形状与类别不相关的数据集（图 3.12），发现受试者的 FMRI 数据中，仍含有类别信息。说明人类在感知物体时，类别信息与形状信息是相互独立的。Coggan 等（Coggan et al.2016）也撰文报道，腹部通道会自动产生对物体类别的选择性。

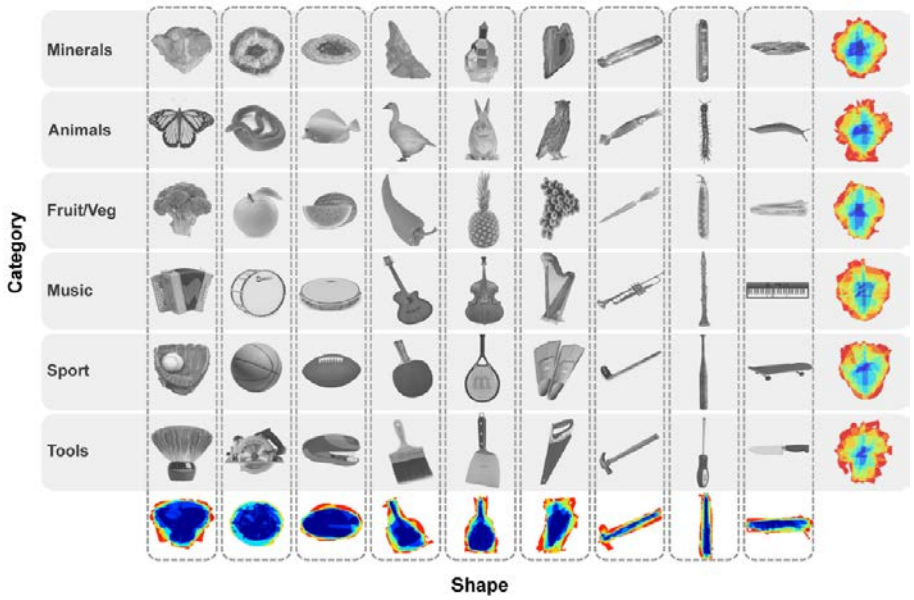


图 3.12：类别与形状独立的测试数据集（摘自 Bracci & Op de Beeck 2016）

Peelen & Downing (2017) 认为，人类视觉皮层的类别选择性，不仅仅用于物体识别。因为物体类别选择性在导航、社会认知（social cognition）和工具使用等同样扮演重要的角色。

人类感知物体时，类别信息起到了主导作用，如苹果、面孔、工具均是基于类别的概念。因此，人类在比较物体的相似性时，首先将同类物体视为“更相似”。Kriegeskorte 课题组（Mue et al. 2013; Jozwik et al. 2016;2017,）对人类对物体相似性判断与人类和猴子 IT 区的物体表达、一种基于类别的感知模型，、一种基于部件的感知模型（如眼睛，耳朵等）以及 DCNN（AlexNet，VGG16）的物体表达进行了比较，发现对图 3.13 的数据集，只有基于类别的感知模型更接近人类对物体相似性的判断。人类 IT 区物体表达对应的 RDM 与人类对物体相似性判断对应的 RDM 也具有比较高的相关性。通过 MDS 降维分析，发现人类判断数据更具有类别聚类的性能，如图 3.14 所示。

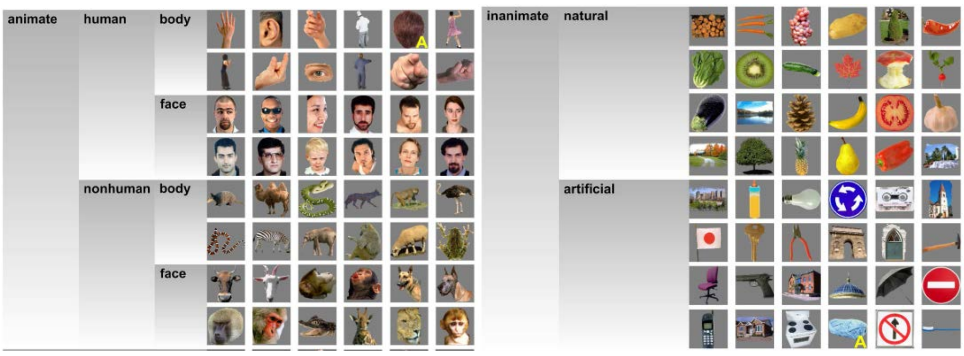


图 3.13： 用于度量人类对物体相似性的主观判断以及人类主观判断与不同模型比较的数据

集（摘自 Mur et al. 2013）

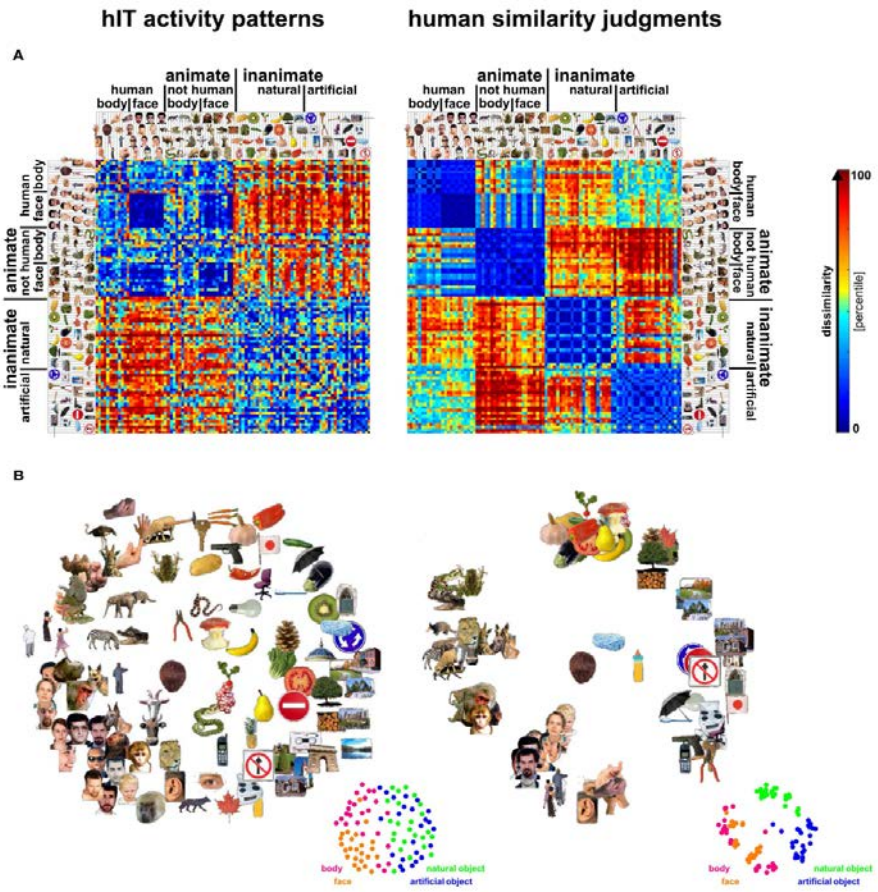


图 3. 14:不论用 RDM 的相似性还是 MDS 降维分析，均表明人类对物体的相似性判断结果包含更显著的物体类别信息（摘自 Mur et al. 2013）

结束语

用深度卷积网络进行生物视觉建模是当前生物领域和类脑研究领域的一个重要研究方向。但笔者觉得，基于深度网络对生物视觉进行建模，由于深度网络有众多的网络结构（architecture）的超参数（hyperparameters）需要确定，而不同的超参数下的网络往往会得到不同的物体表达形式，所以笔者觉得，似乎确定合适的超参数与建模本身的难度没有多大差别。这个方向未来到底有多大前途，需要认真思考，不可盲目跟随，过分乐观估计。

参考文献

Bracci S & a Op de Beeck H. (2016). Dissociations and Associations between Shape and Category Representations in the Two Visual Pathways, The Journal of Neuroscience 36(2):432– 444.

Cadieu et al.（2014）. Deep neural networks rival the representation of primate IT cortex for core visual

object recognition，PLOS Computational Biology 10(12): e1003963

Chang L & Tsao D. Y (2017). The Code for Facial Identity in the Primate Brain, Cell 169: 1013–1028.

Cichy R. M et al.(2015). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence，Scientific Reports | 6:27755 | DOI: 10.1038/srep27755

Coggan D.D et al.(2016). Category-selective patterns of neural response in the ventral visual pathway in the absence of categorical information, NeuroImage 135(2016):107-114

DiCarlo J. J et al (2012). How Does the Brain Solve Visual Object Recognition? Perspective， Neuron 73：415-434，2012.

Dong Q. L et al. (2018). Commentary: Using goal-driven deep learning models to understand sensory cortex, Frontiers in Computational Neuroscience, 19/ doi: 10.3389/fncom.2018.00004

Hong H .et al (2016). Explicit Information for category-orthogonal object properties increases along the ventral stream, Nature Neuroscience 19(4): pp.613-622.

Jozwik K.M et al. (2016). Visual features as stepping stones toward semantics: explaining object similarity in IT and perception with non-negative least squares. Neuropsychologia 83, 201–226.

Jozwik K.M et al. (2017). Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments, Frontiers in Psychology, Vol.8:Article 1726.

Khaligh-Razavi S. M. and Kriegeskorte N. (2014). Deep supervised, but not unsupervised, models may explain IT cortex representation, PLOS Computational Biology10(11): e1003915.

Kheradpisheh S. Z et al.(2016). Deep Networks Can Resemble Human Feed-forward Vision in Invariant Object Recognition, Scientific Reports, 6(32672):1-24.

Kietzmann T. C. et al.(2017). Deep Neural Networks in Computational Neuroscience, bioRxiv preprint first posted online May. 4, 2017; doi: http://dx.doi.org/10.1101/133504

Kourtzi Z & Connor C. E.(2011). Neural representations for object perception: Structure, category and adaptive coding, Annual Review of Neuroscience 34:45-67.

Kriegeskorte et al.(2008). Matching categorical object representations in inferior temporal cortex of man and monkey, Neuron 60:1126-1141.

Kriegeskorte N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing, Annual Review of Vision Science 1:1-22.

Majaj et al. (2015). Simple Learned Weighted Sums of Inferior Temporal Neuronal Firing Rates Accurately Predict Human Core Object Recognition Performance, The Journal of Neuroscience, 35(39):13402–13418

Morcos A. S. et al (2018). On the importance of single directions for generalization, ICLR2018 Best paper.

Mur M et al (2013). Human object-similarity judgments reflect and transcend the primate-IT object

representation, Frontiers in Psychology, Vol. 4: Article 128.

Nili H. et al. (2012). A toolbox for representational similarity analysis, PLOS Biology 10(4):e1003553

Peelena M. V & Downing P. E (2017). Category selectivity in human visual cortex: Beyond visual object recognition, Neuropsychologia 105:177–183.

Rajalingham R et al. (2018).. Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-art deep artificial neural networks. The Journal of Neuroscience doi: 10.1523/jneurosci.0388-18.2018

Yamins D. L. K and DiCarlo J. J (2016). Using goal-driven deep learning models to understand sensory cortex, Nature Neuroscience, Vol.19, No.3, pp.356-365.

Yamins D. L.K et al (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex, Proc. Natl. Acad. Sci. US :111(23):8619-8624.

Zeiler M.D, Fergus R (2013) Visualizing and Understanding Convolutional Networks. ArXiv.org, arXiv: 1311.2901[cs.CV]