课程： 计算机视觉

（第二章）

教 研 室： 脑科学与智能技术

教 师： 胡占义

编写时间： 2017-12-1

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》课程讲义的第二章，旨在对灵长类动物的视觉加工通道进行简单介绍。正像第一章所述，计算机视觉主要研究视觉感知（visual perception），所以计算机视觉研究人员需要对生物视觉有一些基本了解。另外，当前很多人认为深度学习成功的原因在于“模仿了人类视觉系统的信息加工机理”，所以对从事深度学习的研究人员和学生，了解一些生物视觉的信息加工机理也是非常必要的。《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

第二章： 生物视觉中的物体识别通道简介

摘要

本章对生物视觉系统，特别是用于物体识别的视觉腹部通道进行了简单介绍。另外，对群体神经元对物体的编码机制和不同皮层之间的反馈机理也进行了简介。由于读者的对象是从事计算机视觉的研究人员，所以，本章内容尽量简洁易懂，并尽量配以图示，以期对读者能提供一些帮助。

在介绍正文之前，这里首先声明一下，由于有些图是从网上查阅得到的，很难给出具体出处，所以

文中有些图并没有注明具体文献出处，对此笔者表示抱歉并对原作者一并给予致谢。

2.1 为什么要了解一点生物视觉

从计算机视觉当前的研究状况看，以下二方面最有可能取得突破性进展：（１）新的数学方法，如Boykov 等于 1999 年提出的关于图割（GC: Graph Cuts）的优化近似算法（ Boykov et al., 1999）：α-expansion，$(\alpha-\beta)$ −Swap，由于显著提高了优化速度，使得这些算法已成为目前图优化的基本算法；（2）基于生物视觉（或生物视觉启发）的方法。近年来，深度学习在图像物体识别方面取得了“变革性”进展，其层次化的网络结构，神经元“感受野”的逐层增加机制，“简单细胞”，“复杂细胞”等概念，都来自于视神经科学。所以，了解一些生物视觉的知识，特别是群体神经元对图像物体的表达机理（或编码机制），将有助于更好地理解深度网络对图像物体的内在表达机制，并进一步提高深度网络和深度学习在物体识别和场景解释方面的性能。

不同生物种类的视觉系统千差万别，如猴子和老鼠。即使同一种类，如猴子，不同实验手段下得到的一些定量结果也会相差很大。如“猴子的视觉皮层”的总面积到底有多大，不同的文献出处，差距也会很大。所以，对从事信息科学的人员来说，由于习惯于“数字是上帝”，往往会因为神经领域的“数字混乱”感到不知所措。所以，当信息领域人员遇到统计数据差异很大时，应该认为这是“神经领域”的正常现象。神经领域是一个典型的小样本下的外推（extrapolation）领域，统计数据出现大的差异也是不可避免的。

另外，生物视觉既“能力超然”又“愚蠢不堪”。人们可以瞥一眼一张开会的照片，就能给出合理的解释和描述，但人们也经常出现幻觉（illusion）。 所以，从学术的观点探究生物视觉的奥秘是必要的，并在可能的情况下加以借鉴。但如果一定认为生物视觉是计算机视觉“学习的榜样”，也许是完全错误的。人工智能先驱 Marvin Minsky (Minsky, 1986)曾指出智能来源于“多样性”，不是一些特有的“原理”“ What magical trick makes us intelligent? The trick is that there is no trick. The power of intelligence stems from our vast diversity, not from single, perfect principle”. 智能如此，视觉亦如此。当将来有一天人们对生物视觉系统对信息的加工机理基本了解清楚后（可能永远无法完全了解清楚），也许人们会发现，生物视觉系统“没有什么特殊性”，仅仅是“大量神经元连接的一个超复杂网络系统”，这个“复杂系统”仅仅对“生物日常生存需要的一些功能”体现出优良的性能而已，对生物其它不太需求的功能，其“视觉能力”并没有什么优势。所以，人们绝不能过高估计生物视觉的能力，也不宜借鉴生物视觉不太擅长的能力。生物视觉真的有处理图像大数据的优势吗？值得思考。

2.2 猴子视觉系统的基本组成

视觉系统主要由皮层（cortex）和皮下组织（subcortex）构成。认知功能主要靠皮层完成。视觉皮层是指主要对视觉刺激产生响应的皮层。 严格地说，很少有某一皮层区域（cortical area）仅仅对一种刺激产生响应，文献中说的“视觉皮层”， “听觉皮层”和“触觉皮层”等，仅仅表示该区域对某种刺激主要产生响应而已。另外，“视觉功能”必然有视觉皮层参与，但绝不是完全靠视觉皮层完成。视觉认知（visual cognition）的大多数任务都几乎需要大脑的各个皮层区域及皮下组织的协作来共同完成。

如图 2.1 所以，大脑皮层（1.5 -2.0 mm 厚，从上到下又分为 I-VI 的 6 层结构（layer））主要分为 4 个不同的区域：枕叶（occipital lobe），颞叶(temporal lobe)，顶叶（parietal lobe ）和 额叶（frontal lobe）. 由于猴子的大脑皮层与人类相近，所以这里不对它们加以区别。

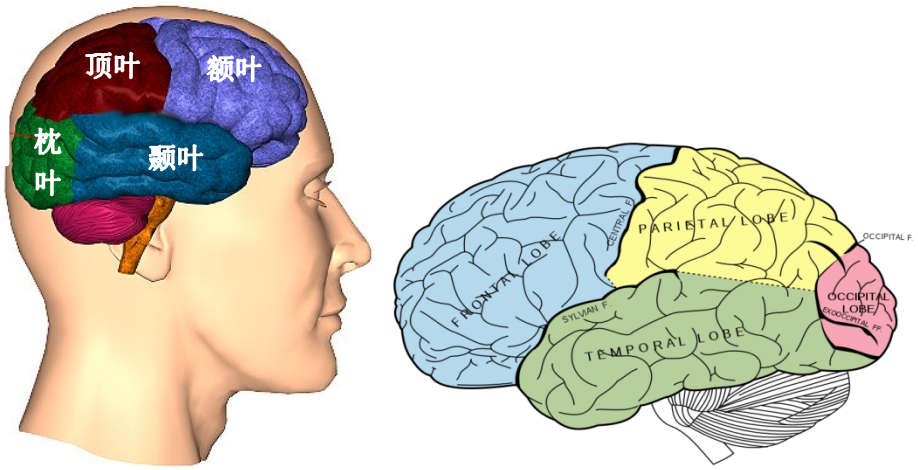

图 2.1： 大脑皮层分为：枕叶、顶叶、颞叶和额叶 4 大区域

视觉信息加工机制，主体上说，是一个层次化加工过程。目前很多文献中认为深度网络（DNN）借鉴了生物视觉加工机制，也主要指这种“层次化结构的相似性”。物体在视网膜（retina）成像后，首先传到后面的枕页，进行底层特征提取。然后传到颞叶，进行物体识别；从枕叶另一路传到顶叶，进行运动检测和空间位置确定等。额叶主要进行高级认知功能。额叶人类快到 30 岁才能成熟，尔后体积慢慢缩小，老年痴呆病人的额叶体积会明显减小。如图 2.2 所以，大脑皮层又可以粗略地分为不同的功能区域，图中不同颜色的颜色区域具有不同的功能：

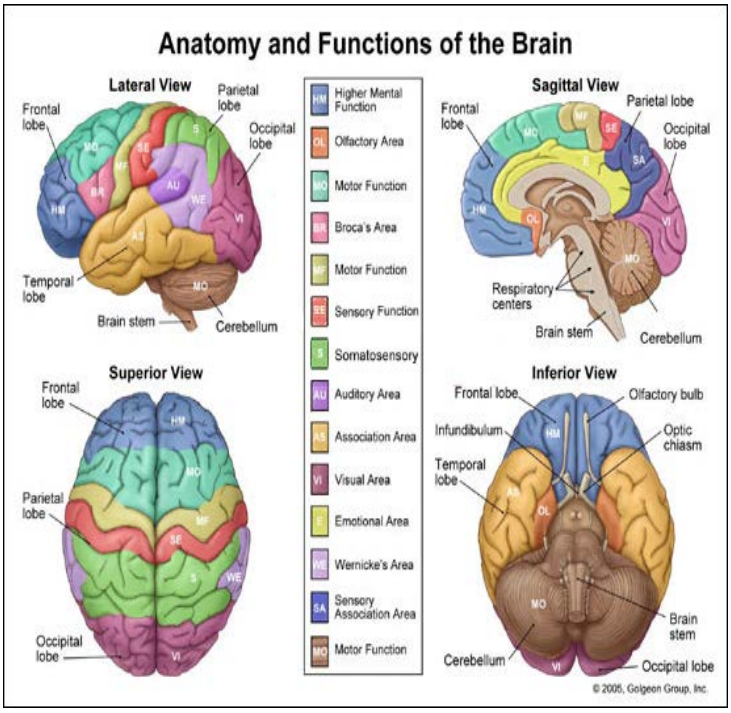

图 2.2：大脑皮层可以分为不同的功能区:左上角为侧视图；右上角为：矢状面 （即从左右脑中间切开的剖面图)；左下角为顶视图；右下角为下观图（即从下面向上看的视图）

物体视觉和空间视觉通道

如图 2.3 所示，视觉信息从视网膜成像并加工处理后，先通过丘脑（ thulamus）的外漆体（ LGN: Lateral geniculate nucleus of thulamus）中继到枕叶的初级视皮层（V1 或 striate cortex）进行初步加工，然后主要分成两个通道（pathway）进行处理（图 2.4）：腹部通道（ventral pathway）进行物体视觉（object vision），背部通道（dorsal pathway ）进行空间视觉（spatial vision）。

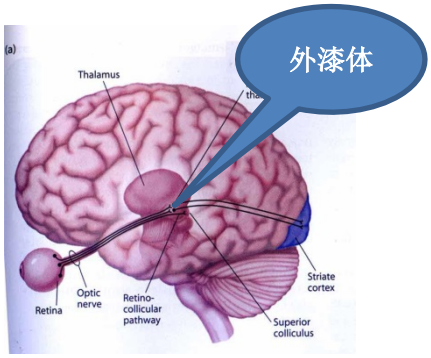

图 2.3：外漆体主要起从视网膜到视皮层之间信号的中继作用

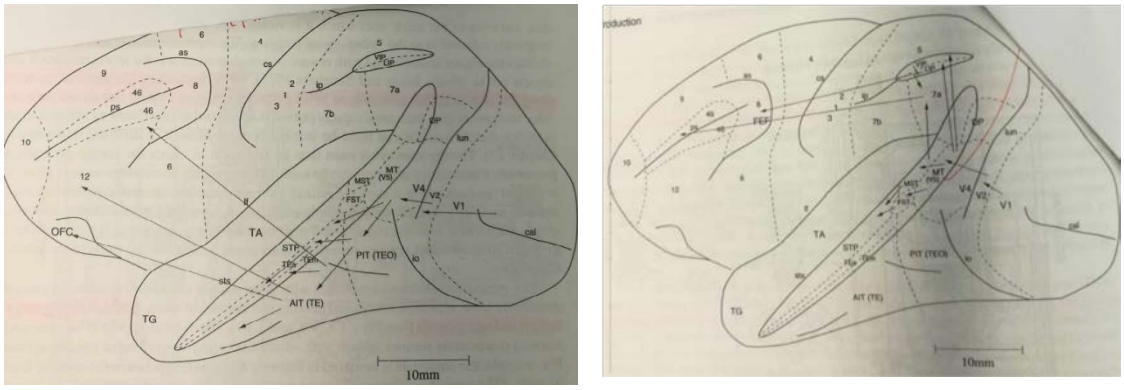

图 2.4：V1 区初加工后的信息分腹部和背部通道进一步逐层处理. (a): 腹部通道(左图) V1→V2→V4→PIT→AIT; (b) 背部通道（右图）V1→V2→MST→LIP→VIP (Rolls & Deco 2004,P.17,P.20)

腹部通道-物体视觉的层次化信息加工通道

腹部通道的主要功能是物体视觉，即对图像物体进行精细识别。图像物体识别是一个层次化处理过程，从 V1→V2→V4→PIT→AIT 区，如图 2.5 所示，神经元对图像刺激形状的选择性越来越复杂。 IT 区（inferior temporal cortex）被认为是视觉物体表达和识别的最高区。

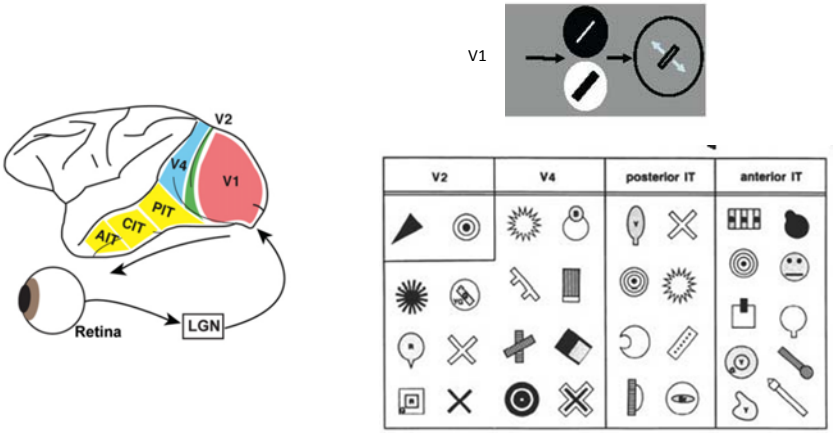

图 2.5：从 V1→V2→V4→PIT→AIT 区，神经元对图像刺激形状的选择性越来越复杂

如图 2.6 所示（*Kravitz D. J. et al., 2012)*，腹部通道的前端（AIT： Anterior IT）又进一步投射到三个皮层区域和三个皮下区域，进行进一步的“高层语义”加工。

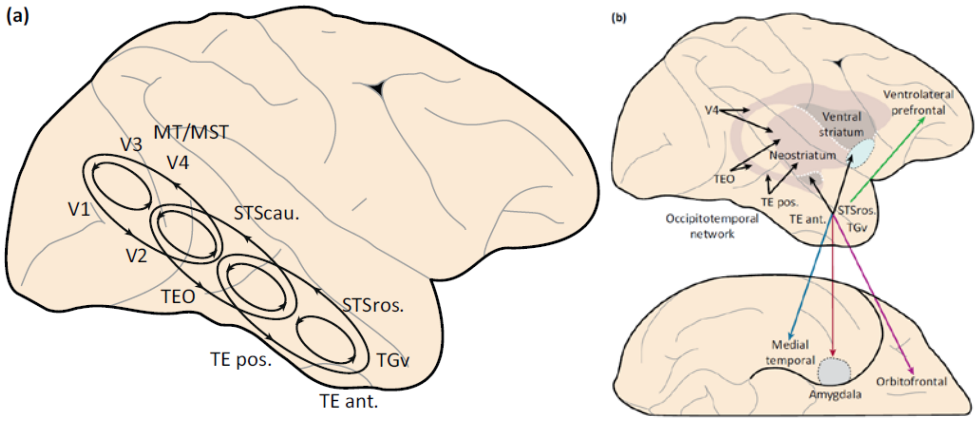

图 2.6：视觉信息从 AIT 区进一步投射到 3 个皮下和 3 个皮层区域(*Kravitz D. J. et al., 2012*). 左图为物体识别的基本通道，右图为与高级认知相关的脑区的信息传递关系

AIT 区进一步投射到的三个皮下区域及功能为：

到新纹状体（neostriatum），支撑依据视觉的习惯和技能形成（support visually-dependent habit formation and skill learning）；

到腹侧纹状体（ventral striatum），支撑刺激“效价”的赋值（support the assignment of stimulus valence）。 效价（Valence） 是一个心理学术语，指根据“吸引力”（attractiveness）赋予某人或某事的评价值。

到杏仁体（amygdala），支撑情感刺激处理（supports the processing of emotional stimuli）

AIT 区进一步投射到的三个皮层区域及功能为：

到内侧颞叶（ medial temporal ），支撑物体长时和上下文记忆（supports long term object and object-context memory）

到眼窝前额皮质(orbitofrontal)，参与奖励处理(mediate reward processing);

到腹外侧前额叶（ventrolateral prefrontal），参与物体工作记忆（mediate object working memory）。

背部通道—空间视觉的层次化信息加工通道

空间视觉指为“行为和动作”服务的视觉， 如视差计算，运动估计，坐标系变化等。空间视觉主要在枕叶和顶叶（parietal lobe）完成。如图 2.7 所示（*Kravitz D. J. et al., 2011)*，从 V4→ 顶页后，处理后的信息从后顶叶皮层 （Posterior parietal：PP）又投射到三个不同的脑皮层区。具体情况如下：

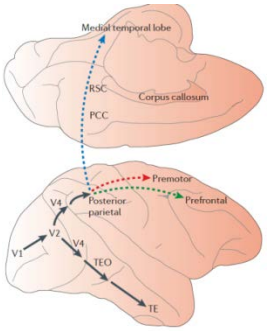

图 2.7: 背部通道视觉信息从后顶叶皮层进一步投射到其它 3 个皮层区域（*Kravitz D. J. et al., 2011)*

投射到前额皮层(Prefrontal)，参与空间工作记忆（Spatial working memory）

投射到运动前皮层（Premotor），执行视觉引导的动作（Visually guided action）

投射到内侧颞叶（Medial temporal lobe），用于导航（Navigation）

腹部通道从 IT 区（背部通道从 PP 区）后，处理的信息不再为单纯的视觉信息处理，视觉信息参与后续的多模态信息融合、记忆和推理等高级认知功能。所示，视觉能力绝不是仅仅靠“视皮层”完成的。视觉能力是大脑各功能模块协同处理的结果。所以，“理解视觉系统”与“理解整个大脑”本质上没有多少区别。

视网膜（retina）及其信息加工机制

不少计算机视觉研究人员，往往认为视网膜与照相机一样，是一种光电转换机构。 事实上视网膜有 3 层信息加工过程。 如图 2.8 所示，感光细胞（photoreceptor）进行光电转换后，信号通过双极细胞层（bipolar cells）中继后，通过神经节细胞层（ganglion cells）进一步处理，最后输出到丘脑的外漆体 (LGN)。从图 2.8 可以看出，视网膜的信号为“从后向前”的处理机制。

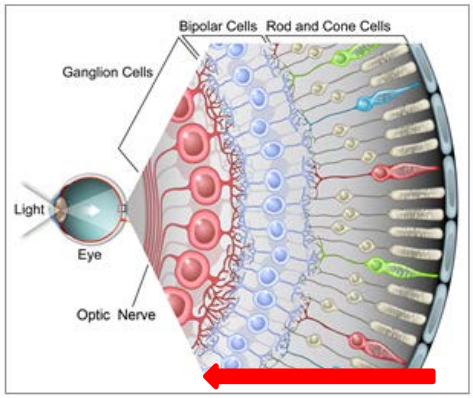

图 2.8： 视网膜从感光细胞到双极细胞再到神经节细胞从后向前的三层处理结构

感光细胞（photoreceptors）

感光细胞进行光电转换，分为二类：锥状细胞（cones）和杆状细胞（rods）。人类视网膜大约有 5-6 百万个锥状细胞，1 亿 2 千多万个杆状细胞。杆状细胞主要用作夜视。杆状细胞可以捕获单个光子，多个杆状细胞共同作用，在微暗的光度下形成夜视能力。锥状细胞主要分红、绿、蓝三种类型，它们之间的比例约为：红：62%；绿：32%，蓝：2%。这种比例关系与目前的彩色相机中不同颜色的 CCD 单元个数相同的结构也有大的区别。锥状细胞主要分布在视网膜的中央凹（fovea）周围，是产生正常清晰视觉的神经基础。由于中央凹非常小，所以正常情况下人们要不断地调整“注视点”（Focus of attention）进行大范围观察。 杆状细胞和锥状细胞个数随着视张角（偏离中央凹的偏离角）的分布如图 2.9 所示：

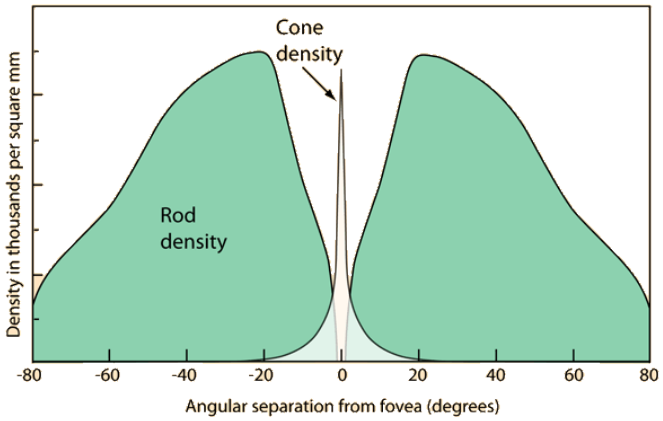

图 2.9： 锥状细胞主要分布在中央凹，杆状细胞密集分布在整个视网膜

双极细胞和神经节细胞

双极细胞（bipolar cell）主要起信号中继作用。神经节细胞（RGC: Retina Ganglion Cell）大约有 1 百多万个。分为大细胞（Paraso RGC）和小细胞（Midget RGC）. 大细胞对运动敏感，对空间频率不敏感，反应速度快；小细胞对颜色敏感，对空间频率敏感，反应速度慢。大细胞与小细胞的比例个数约为：1：9。从上面的介绍可知，从感光细胞到神经节细胞，细胞个数得到大量压缩，视觉信息已得到很复杂的处理，所以，视网膜不能等价为一台照相机。我们日常感受到的五彩缤纷的世界，都是从一百多万个神经节细胞输出的信息进一步加工得到的。从这个意义上说，提高照相机的分辨率未必是提高“计算机视觉”能力的有效途径。

细胞的感受野（RF: receptive field)

细胞的感受野是指该细胞对空间视张角内（或视网膜对应区域内）刺激产生响应的区域，图 2.10 为二个不同神经元的感受野示意图：

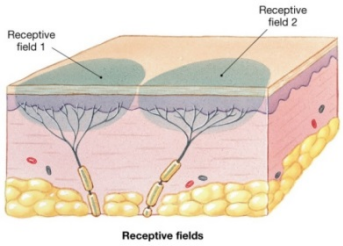

图 2.10：细胞的感受野

当一个神经元接收多个神经元的输入时，该神经元对应的感受野为多个输入神经元感受野的组合，如图 2.11 所示：

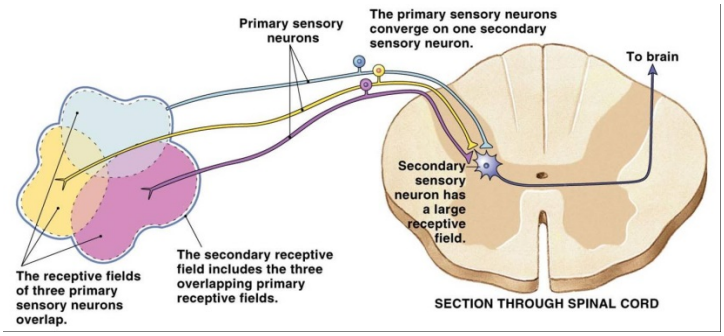

图 2.11：多个输入细胞感受野的集合构成接收细胞的感受野

从 V1→V2→V4→PIT→IT 信息的逐层处理过程中，对应的神经元的感受野越来越大，每层之间感受野增大的系数大体为 2.5. V1：1.3 度；V2: 3.2 度; V4: 8.0 度; PIT:20 度; IT 50 度。感受野越大，说明对应神经元编码的图像特征越复杂 （ Rolls & Deco 2004. P.254）

2.3 腹部通道和物体识别

腹部通道是物体表达和识别的通道. 深度网络所谓的“模拟脑结构”也主要指腹部通道的这种分层处理机制。首先，生物视觉系统的层数（解剖区）一般也就是 10 多层，不像 DNN 动咋几十层，甚至几百层，上千层；其次，生物视觉同一层内部的神经元连接更加普遍（约占 80%），层与层之间的连接约占 20%；另外，生物视觉的跨层连接非常普遍，基本上很少有不同区域没有神经连接的现象，仅仅体现为连接强度的差别而已。这些均与目前的 DNN 结构有区别。

图 12 给出了皮层不同区域之间的连接关系及连接强度。图中对应的方块大小表示皮层的面积大小，连线粗细表示连接强度。

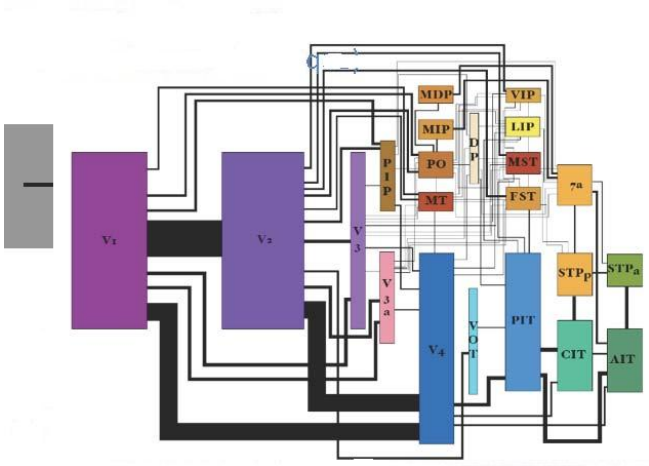

图 12： 视皮层不同区域的面积大小及连接强度

图 13 给出了每层神经元的个数、输入输出神经元的个数和从刺激出现到该层神经元出现响应的潜伏期（latency）.例如，从视网膜到达 V1 区的潜伏期约为 50 ms, 达到 AIT 的时间约为 100 ms. V1 区接收 LGN 的输入个数约为 1 百万左右，V1 区的神经元个数约为 1 亿 9 千万，V1 区的输出约为 7 千 7 百万（DiCarlo J. J. et al., 2012）。图中其它区域的数字含义相同。

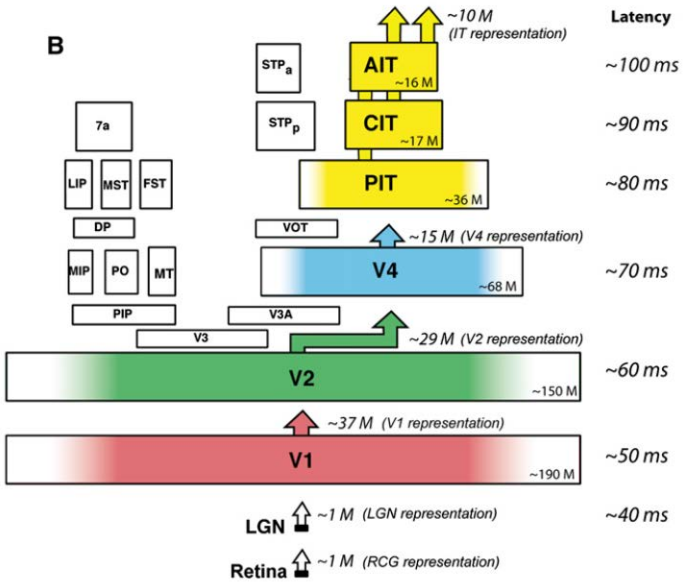

图 13：腹部通道不同区域的输入、输出和区域内部神经元的个数分布（DiCarlo J. J. et al., 2012）

目前除了 V1 区细胞对刺激的加工机制比较清楚外， V2， V4， IT 区的功能尚不是很清楚。大体上说，从 V2 区开始神经元对错觉轮廓（illusory contour）有反应，已形成色调敏感区；V4 区为颜色感知的主要区域，参与曲率计算、运动方向选择和背景分离 ；IT 区为物体表达和识别区。这里需要指出

的是， IT 区并没有形成完全的与视角无关的表达。研究表明（Lafwe-Sousa & Conway，2013）， IT 区细胞的感受野的大小分布不均匀。既有大感受野细胞，也有中小感受野细胞。IT 区存在小感受野细胞，说明这些细胞仅仅对物体的局部敏感。图 14 为 IT 细胞的感受野分布图：

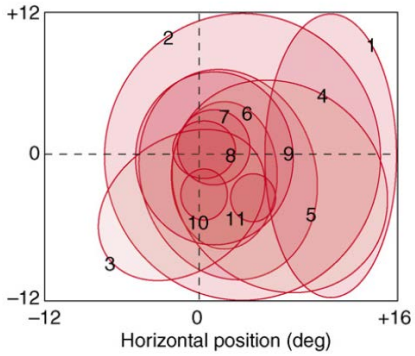

图 14：IT 区同时包含大感受野和中小感受野的神经元（感受野大小用椭圆大小表示，Lafwe-Sousa & Conway，2013）

简单细胞和复杂细胞 （simple cell and complex cell ）

简单细胞和复杂细胞是 Hubel 等人研究猫的 V1 区细胞对图像边缘（和运动）的响应时提出的感念（ Hubel & Wiesel, 1959）. 简单细胞指其响应对边缘（运动）的具体位置密切相关的细胞，Hubel 等给出的简单细胞的感受野结构如图 15 所示. 复杂细胞，指对边缘（运动）在一个小范围内变动其响应不明显变化的细胞，Hubel 等认为复杂细胞是由简单细胞组合得到的，则复杂细胞的感受野就是简单细胞的感受野的空间组合，如图 2.11 所示。Hubel 等的复杂细胞，并不是指“对复杂刺激有响应的细胞”。目前在深度学习领域，“复杂细胞”的概念有点“过度泛化”。在视觉通道的高层区域，如 IT 区，IT 神经元对输入刺激的形状选择已很复杂，但似乎没有人称 IT 细胞为复杂细胞。神经领域有“老祖母细胞”(grandmother cell)的提法，是指该细胞对特定面孔有响应的细胞。目前看来，是否存在老祖母细胞也尚无定论。因为实验时发现，在上百张面孔图像中，有的细胞确实仅对其中的一幅图像有响应，但当测试图像多时，特别是当图像之间的相似性增大时，那些老祖母细胞是否仍仅对一幅特定面孔图像有响应、对相似性面孔图像无响应仍是一个未定的问题。

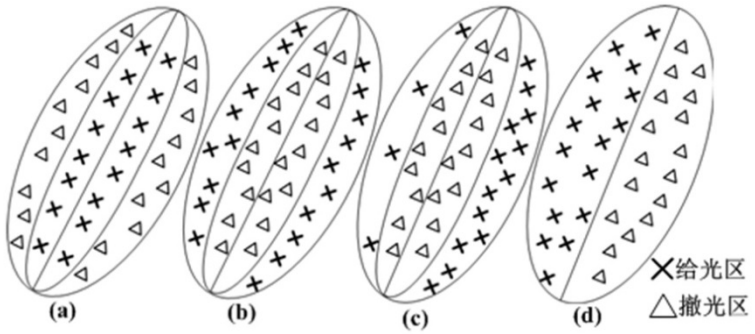

图 15： 几种简单细胞的感受野（Hubel & Wiesel 1959）

神经元的响应表达和物体表达

单个神经元的响应就是指神经元对视觉刺激发放的脉冲序列。目前的研究表明，将一段时间（如 IT 神经元在 50ms -180ms）神经元发放脉冲个数的平均数作为该神经元对刺激的响应，是一种有效的响应表达方式。所以，文献中关于神经元的响应，一般就是指“给定时间段内发放脉冲个数的平均值”。

单个神经元的作用很小，群体神经元才能对视觉物体进行表达（或编码）。群体神经元对某个视觉物体的表达，就是将同一皮层区域多个神经元对该物体的响应表示为一个向量，$\mathrm{X} = (x_1 \quad x_2 \quad \cdots \quad x_n)$, 其中 x_i 为第 *i* 个神经元对输入图像刺激的响应，用该向量来作为对该图像物体的表达。

这里需要强调一下，神经文献中在比较不同表达模型对同一物体的表达之间的关系时，如果二种表达之间满足一个线性变换，则认为是相同表达。例如，假定猴子 IT 区对第 *i* 个输入图像物体的表达为 X_i^{Monkey}， 该物体在某种深度网络下的表达为 X_i^{DNN}, 如果这二种表达之间满足线性变换关系，即：$X_i^{\mathrm{DNN}} = \boldsymbol{A}X_i^{Monkey} + \boldsymbol{B} \ \forall\, i$， 则认为 X_i^{DNN} 和 X_i^{Monkey} 为相同表达，或深度网络给出了一种优秀的对猴子 IT 区的物体表达的定量预测模型。神经元个数固定后，由于输入图像的个数可以非常大，所以对二种不同的表达，上面的线性变换关系不可能对所有输入图像都满足。当然具体应用中，任何时候不可能严格满足这种线性关系，需要利用某些统计方法来进行显著性验证。

群体神经元对图像物体的表达方式

对视差（disparity）和运动（motion）等刺激，由于这些刺激可以用参数描述，如运动可以用运动方向和大小来描述，所以神经元对这些刺激的响应曲线可以比较容易测定和参数化表示。但对图像物体，很难给出参数化的响应曲线。文献中分析猴子对图像物体的表达时，先将某个神经元对所有刺激的响应表示为一个概率密度函数，称为该神经元的选择性函数，或将所有神经元对同一刺激的响应（即群体编码）表示为一个概率密度函数，称为群体神经元的稀疏性函数，然后对所有选择性函数（或稀疏性函数）进行统计分析。如利用分布函数的 4-阶矩来刻画其非高斯性，用尾巴指数（Tail index）来刻画分布函数

的大尾巴性等。分布函数的“尾巴”越大，说明个体神经元的选择性越高，或群体神经元的编码稀疏性越高。

Lehky 等（Lehky R. S. et al., 2011,2014）对二只猴子 674 个 IT 神经元对 806 幅图像的响应（选择性函数和稀疏性概率密度函数）分析表明，IT 神经元的编码具有如下特性：

（1)：：单个神经元的选择性都不高，但 IT 区存在大量这样的神经元。这种 IT 区神经元的编码方式与传统模式识别理论的物体表达方式不同，主要体现在神经元的表达是一种“过表达”（over-complete representation）,而传统模式识别领域的物体表达方式则希望不同的表达尽量相互独立。IT 神经元的这种过表达方式与目前 DNN 神经元的表达方式具有相似性；

（2)：尽管 IT 区有大量神经元，如图 13 所示，AIT 区有约 6 千 6 百多万个神经元，但这些神经元响应之间独立的个数不多，一般不超过 100 个。

目前对猴子的物体视觉系统研究表明 (Yamins K. L. D. et al., 2014, 2016a), 如果一种数学表达方法可以与猴子 IT 区神经元对物体的表达相媲美，则该物体表达方法对输入图像物体一定同时具有高的物体分类（categorization）和鉴别（identification）能力，即这种表达是一种优秀的表达。

董秋雷等的近期工作表明 (Dong Q .L. et al., 2017)，DNN （以 VGG19 作为 DNN 研究模型）神经元的编码特性与猴子 IT 神经元的编码特性有本质的差别。他们的结果同时表明， 一种优秀的表达不一定非要与猴子 IT 神经元的表达相似。

猴子视觉系统对脸孔图像的表达

鉴于面孔在生物生存和社交中的重要性，面孔识别一直是神经领域研究猴子物体识别机理的重要内容。2010 年，Freiwald 等（Freiwald &Tsao 2010 ）发现，猴子对面孔的识别逐渐从视角相关到视角无关。如图 16 所示，猴子后外侧皮层（PL: posterior lateral cortex ）的神经元的响应与视角相关，但前内侧皮层（AM: Anterior medial cortex）的神经元的响应已与视角无关，达到了对面孔图像的不变识别（invariance）。

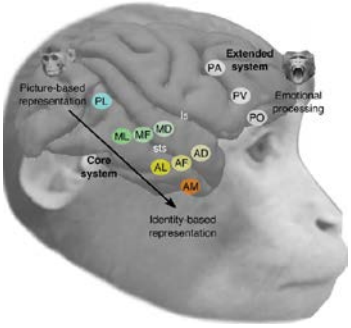

图 16：从 PL 区到 AM 区，神经元对面孔的响应从与视角相关到无关（Freiwald &Tsao 2010 ）

Landi 等（Landi & Freiwald 2017）进一步研究发现，猴子除了上面面孔识别的核心系统（core system）外，另有二个区域: 一个位于颞极区（temporal pole），另一个位于鼻周皮层区（perirhinal cortex），即图 17 所示的 TP 和 PR 区，主要负责对“熟悉面孔”（familiar face）的识别。而且，对熟悉面孔的响应呈非线性性，可能体现了人们对曾经见过的人“突然想起来了”的现象。

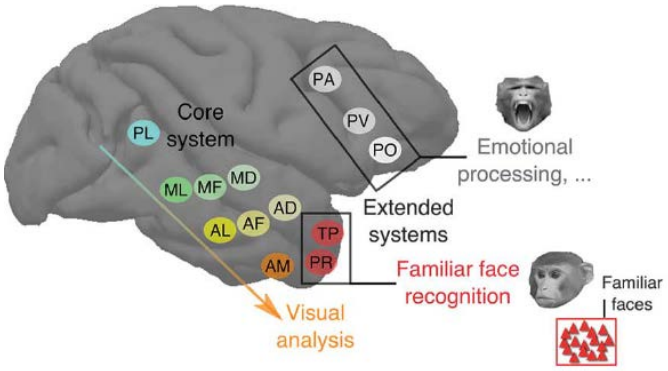

图 16：猴子 2 个新的仅对熟悉面孔识别的区域（Landi & Freiwald 2017）

Chang 等（ Chang& Tsao 2017）研究发现，当面孔图像用一个包含纹理和几何信息的参数向量表示时，则每个 IT 神经元均有自己的一个特定向量（STA：Spike-Triggered Average），该神经元对任何一幅新输入的面孔图像的响应就是该神经元的 STA 与输入图像参数向量的内积。说明猴子 IT 神经元在面孔图像的参数空间具有“简单而优美”的线性编码机制。这一结果与目前的 DeepFace 等深度网络对面孔图像的表达机制有很大区别。

神经元响应定量建模 – 目标驱动的框架

DiCarlo 团队 2016 年在 Nature Neuroscience 发表了一篇“Perspective”文章，提出了所谓的“目标驱动的感知信息建模方法” (Yamins & DiCarlo., 2016a)。Dicarlo 等认为，只要将物体分类性能作为训练一个深度卷积网络（HCNN）的优化目标函数， 则训练好的 HCNN 就能很好地预测猴子 IT 神经元对物体的编码。在这种目标驱动的框架下，他们通过在 ImageNet 上训练了一个 6 层的卷积神经网络(Yamins & Dicarlo 2016b)，训练时代价函数仅仅为物体分类精度，则训练好的网络的输出不仅可以“定量预测”IT 神经元的响应，而且可以预测物体的位置、大小等几何信息。由于训练时仅仅“控制”物体的分类能力，没有使用任何 IT 神经元响应的信息，所以他们称这类建模方法为“目标驱动”的建模。

目标驱动的框架在一些情况下可以对 IT 神经元的响应进行很好预测，但是否如 DiCarlo 等认为的是一种“通用性建模”方法值得磋商。文献中已有报道(Li et al, 2016)，相同的 DNN 深度网（Alex 网络）(Krizhevsky A.et al., 2012)，当使用 4 种不同的初值训练时，所得到的 4 个 Alex 网络的图像正确分类率基

本相同，但 4 个不同 Alex 网最后一层神经元之间的表达无法用一个线性变换表示。作者将这种现象称之为“收敛学习”与“发散学习”现象。这些结果对 DiCarlo 等提出的“目标驱动的感知信息建模方法”提出了挑战。因为网络结构相同、仅仅不同初值下训练得到的不同 DNN，就出现“发散学习”现象，当网络结构不同时，这种“发散学习”现象就会更普遍。另外，从数学的观点看，HCNN 的神经元表达是一种“过表达”,而同一问题的多个等价过表达之间一般不存在线性变换关系。

2.4 猴子 IT 区神经元的物体表达与其它物体表达方式之间的比较

正像前面所说，文献中已有结果表明，如果一种数学表达方法可以与猴子 IT 区神经元对物体的表达相媲美，则该物体表达方法必然是一种优秀的物体表达方法。那么，如何比较评价猴子 IT 区（或其它区）的物体表达性能呢？Dicarlo 组对此进行了研究(Cadieu F. C. et al., 2014)。他们比较依据的基本原理是：

给定同一分类（回归）问题的两种表达，如果在同样分类（预测）精度情况下，一种表达下需要的主成份分量个数少，另一种表达下需要的主成份分量个数多，则对该问题而言，需要主成份分量个数少的表达优于个数多的表达。

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$k_\sigma(*,*)$, 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：

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

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 17 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：

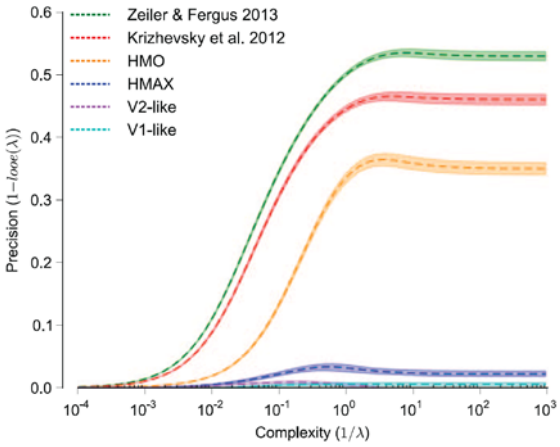

图 17：正则项系数 $\frac{1}{\lambda}$ 可以刻画线性回归问题的复杂度(Cadieu F. C. et al., 2014)

图 18 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：

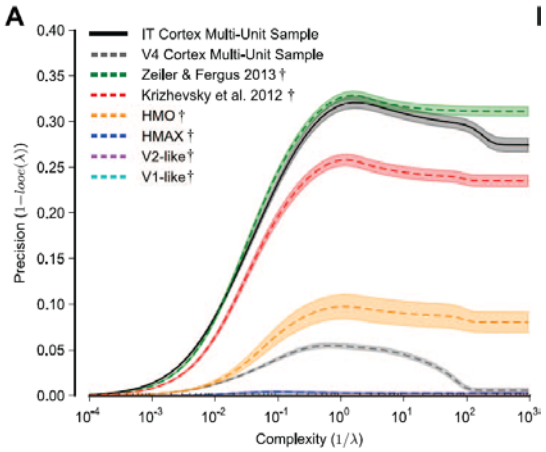

图 18：Zeiler 等的深度网可以与猴子 IT 神经元的物体表达能力相媲美(Cadieu F. C. et al., 2014)

研究与猴子 IT 神经元表达进行比较的最大困难是猴子记录的响应数据少。在小数据下的结果是否“真实反映”了大数据下的性能，目前仍是一个有待进一步研究的课题。

2.5 大脑皮层之间的反馈

视觉皮层之间从高层到底层的反馈机制到底是什么？是从高到低逐层反馈？还是从高层先反馈到 V1 区，然后从 V1 区再逐层上传到 V2 和 V4 等后续区域呢？这个问题长期以来一直是一个有争议的问题，即使目前文献中也仍然存在不同的看法。Buffalo 等（Buffalo E. A. et al., 2010）通过研究注视效果与“潜伏期”长短的关系，发现注视效果到达 V4 区的时间最短，然后是 V2 区，最后是 V1 区，说明“注视”的反馈效果是从高层到底层的，如图 19 所示：

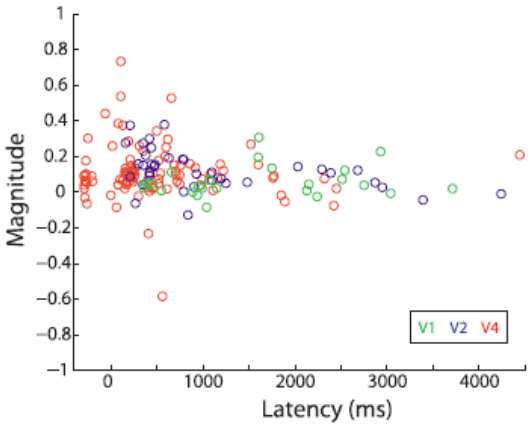

图 19：视觉注视效果是从高层到底层的逐层反馈机制（Buffalo E. A. et al., 2010）

对计算机视觉的研究人员来说，每当提及反馈，首先问的问题是：反馈什么？这似乎是一个“不太科学”的问题。事实上，包括 VGG 等前馈深度神经网络，底层对高层到底上传了什么，人们似乎也并不清楚。文献中给出的一些有特定意义的“感受野”，如对应眼睛、耳朵等的神经元感受野，也仅仅是挑选出来的极少个例，并不具有一般性。所以，对一个复杂的网络研究其反馈机制时，似乎仅仅研究“反馈环路构成”和“对应的反馈强度”就足够了。“反馈强度”由于受“视觉任务”影响，也是一个难以定义和测定的量，所以，计算机视觉界似乎可以借鉴“脑连接组学”（Connectomics ）的成果，特别是不同皮层区域之间的神经元的投射比例（ neuron projection ratio）关系，来探索视觉反馈机制。

Karl Friston （2003）和 Markov 等 （Markov T. N.et al., 2011,2014）的研究表明，反馈具有如下的一些基本属性：

- 高层到底层的反馈起调制作用. “调制”表示“反馈无法单独激活相应神经元”， 仅仅对从底层输入的信号“增强”或“抑制”；调制同时表示“高层对底层作用的区域比较大”。前馈“底层对高层作用的区域（感受野）相对要小”；
- 皮层之间的反馈连接现象比正向连接更普遍；
- 跨层反馈现象更普遍；
- 反馈强度与距离呈指数规律下降；
- 反馈既可以通过双向神经回路（底层到高层和高层到底层），也可以通过单向回路（仅仅高层到底层）进行；
- 反馈的信号较前馈要慢 （50 多毫秒），前馈层与层之间约为 10 毫秒左右；

Markov 等（Markov T. N.et al., 2011,2014）通过测定不同脑皮层区之间的神经元投射比例关系 FLN (Fraction of Labeled neurons) 发现：如果 A 区投射到 B 区的距离为 d， 则投射比例（强度）FLN 的概

率服从下面的指数分布，其中 τ =0.188 mm^{-1}：

$$P(d)= e^{-\tau d}$$

A 区投射到 B 区的神经元比例 $FLN_{A\to B}$ 的定义为：

$$FLN_{A\to B} = \frac{\text{区域 A 投射到区域 B 的神经元个数}}{\text{区域 B 的总神经元个数}}$$

Markov 等发现，给定目标区域，其它区域投射到目标区域的 FLN 服从负对数正态分布（lognormal）。也就是说，其它投射区域到目标区域之间的距离服从正态分布。换句话说，大多数投射区域与目标区域的距离在正态分布的均值附近（ $\mu = 26.57mm$，$\sigma = 10.11$ mm)。

图 20 为其它皮层对视觉皮层 V1, V2, V4 区投射的神经元比例 FLN 的分布图：

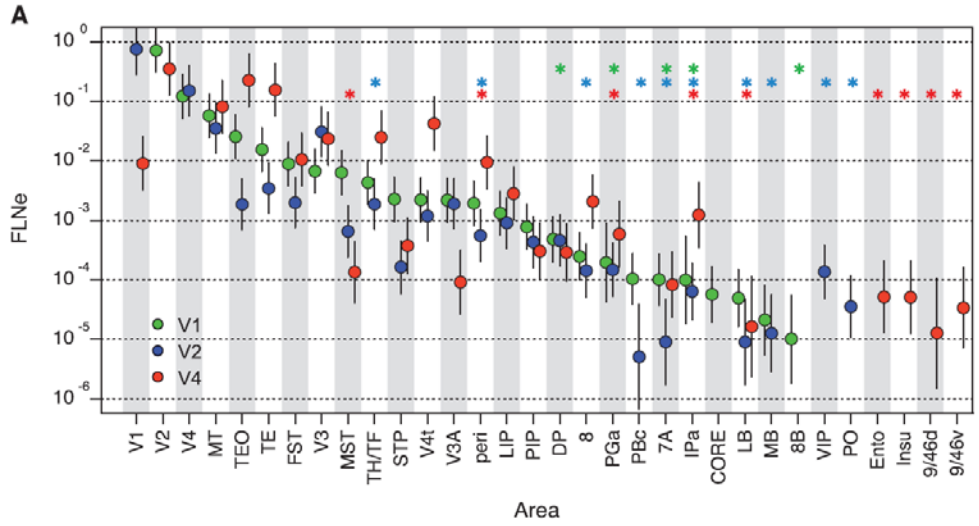

图 20：其它脑皮层对 V1,V2, V4 区投射的神经元比例的分布图（Markov T. N.et al., 2011）

从图 20 可以看出，即使对用于初级加工的视皮层 V1, V2, V4，也有大量其它区域的神经元的投射关系。一般来说，邻域之间的投射要多，但也有一些例外。有些区域尽管不是邻域关系，但它们之间的神经投射也接近邻域投射的比例强度。如果把大脑皮层之间的不同区域看作脑网络的节点的话，那么对应的脑皮层网络是一个“密集连接”的网络，而不是过去人们认为的仅仅存在局部连接的“小世界”（Small world）的网络结构。另外，从图 20 也可以看出，其它皮层投射到 V1, V2, V4 的 FLN 的数值有 5 个数量级的差别。如何借鉴这些信息，来构建具有反馈机制的深度网络，是一个值得探索的未来研究方向。

上面对生物视觉，特别是猴子的腹部物体识别通道，进行了简单介绍，相关内容以期对计算机视觉研究人员有所帮助。另外，空间视觉需要借鉴立体视觉（Stereopsis）和神经视差加工机制，相关内容可参阅孔庆群等的关于视皮层中的神经对视差的加工机理方面的综述文章（孔庆群等，2011）

参考文献

Boykov Y. et al. (1999). Fast approximate energy minimization via graph cuts， ICCV1999；

Cadieu F. C. et al (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition，PLoS Computational Biology 10(12): 1-18;

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

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

Dong Q. L. et al. (2017). Statistics of Visual Responses to Object Stimuli from Primate AIT Neurons to DNN Neurons, Neural Computation.

Kravitz D. J. et al.(2012). The ventral visual pathway: An expanded neural framework for processing of object quality, Trends in Cognitive Sciences 17(1): 26-49

Kravitz D. J. et al. (2011). A new neural framework for visuospatial processing，Nature Review: Neuroscience 12:217-230

Friston K. (2003). Learning and inference in the brain. Neural Networks 16:1325:1352;

Freiwald W A & Tsao D Y(2010). Functional compartmentalization and viewpoint generalization within the macaque face-processing system. Science 330(6005): 845.

Hubel D H & Wiesel T N (1959). Receptive fields of single neurons in the cat's striate cortex. *The Journal of Physiology* **148**(3): 574-591

Krizhevsky A et al (2012). ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems 25. pp.1106–1114.

Lafwe-Sousa R et al.(2013). Parallel, multi-stage processing of colors, faces, and shapes in macaque IT cortex, Nature Neuroscience. 16:1870-1878;

Landi S.M & Freiwald W. A. (2017). Two areas for familiar face recognition in the primate brain, Science 357, 591–595

Lehky R. S. et al. (2011). Statistics of visual responses in primate inferotemporal cortex to object stimuli, J. Neurophysiology 106:1097-1117;

Lehky R. S. et al. (2014). Dimensionality of object representations in monkey inferotemporal cortex, Neural Computation 26, 2135-2162;

Minsky M. (1986). The Society of Mind, New York: Simon & Schuster. ISBN 0-671-60740-5 ,P.308；

Li Y et al. (2016). Convergent learning: Do different neural networks learn the same representations? In International Conference on Learning Representations 2016;

Markov T. N. et al.（2011）. Weight consistency specifies regularities of macaque cortical networks, Cerebral Cortex 21:1255-1272;

Markov T. N. et al. (2014). Anatomy of hierarchy: Feedforward and feedback pathways in Macaque visual

cortex, The Journal of Comparative Neurology: Research in Systems Neuroscience 522:225-259.

Rolls. T. W & Deco G. (2004). Computational Neuroscience of Vision, Oxford University Press;

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

Yamins D. L. K & DiCarlo J. J (2016a). Using goal-driven deep learning models to understand sensory cortex, Nature Neuroscience 19(3):.356-365;

Yamins D. L. K & DiCarlo J. J (2016b). Explicit Information for category-orthogonal object properties increases along the ventral stream, Nature Neuroscience 19(4):613-622;

Zeiler MD & Fergus R (2014). Visualizing and Understanding Convolutional Networks. ECCV2014;

孔庆群等（2011），视皮层中的视差计算，《自动化学报》37(6): 645-657.