09 认知神经系统

北师大教授毕彦超：人类大脑的知识表征

转载自：AI 科技评论

人工智能和认知神经科学都在尝试打开“智能”的黑箱，两者应相互对话、相互帮助，才能共同快速发展。一方面，脑科学能帮助人工智能专家构思出更好的网络结构、更好的算法，从而推动人工智能的发展； 另一方面，我们也经常发现，AI 专家发明出的人工智能算法，经常和生物体处理信息的方式极为类似。

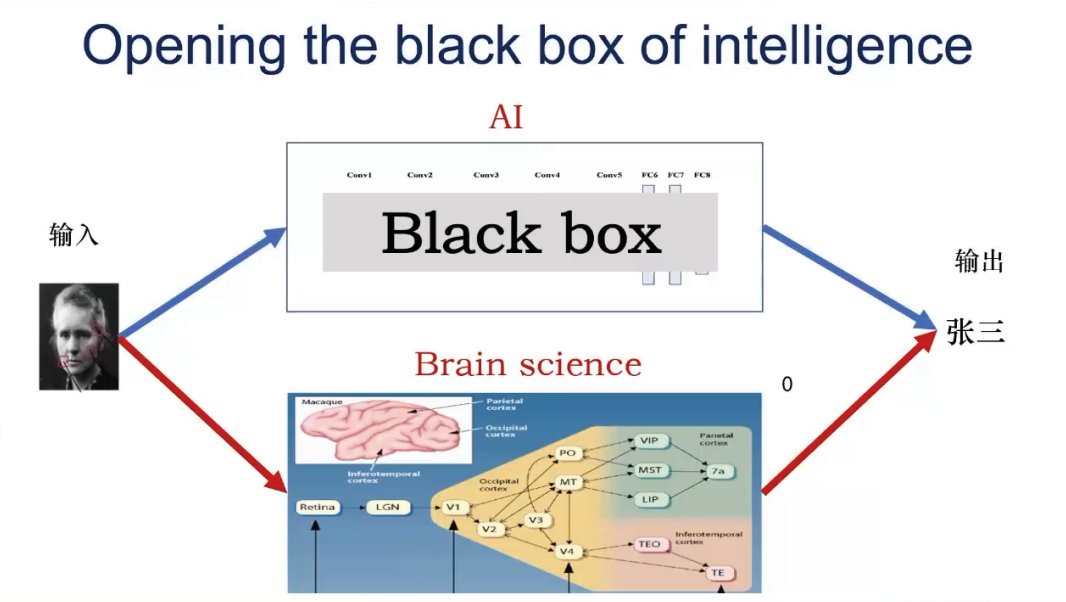


图 1：打开人工智能的黑箱

那么，人工智能发展到最后真的会变得完全和生物大脑一样吗？不一定，因为两者服务于不同的目的。**人工智能要实现的是具有专用功能的机器，而生物智能要实现的是能适应大自然环境的有机体。**

但是生物大脑是亿万年进化的产物，所以它在进化过程中已经摸索出极佳的信息处理的架构和算法，这些架构和算法可以为发展人工智能带来启发。

所以从原则上来说，**两者存在交集，但并没有包含关系**。

6 月 22 日，北京智源大会举行了认知神经基础专题论坛，来自北京师范大学认知神经科学与学习国家重点实验室的毕彦超教授、北京大学心理与认知学院的方方教授、清华大学心理学系的刘嘉教授、北京大学计算机系的吴思教授、中国科学院自动化研究所的余山教授分别做了报告，共同探究认知神经科学能为 AI 带来什么启发。

毕彦超教授做了《人类大脑的知识表征》的报告。毕彦超教授在哈佛大学获得心理学（认知、脑、行为）博士学位，在人脑实现语义知识表达方面做过很深入的研究。

在报告中，毕彦超教授汇报了三个实验，解释了人脑有两套知识表征模式。一套是感觉信号来源的知识编码，另一套是语言信号来源的知识。两套编码系统的信息内容和编码方式都有不同。

以下是演讲全文。

一、知识在大脑哪里

AI 的知识表征一般指从文本提取各种知识图谱，而人脑里其实存在很多非语言描述的知识。

举两个例子，有的大脑损伤的病人，给他一个剪刀，他知道这是剪刀，也知道剪刀是用来剪东西的。但是他完全不知道该怎么用，连应该怎么拿都不知道。

另外一类病人，我们也给他一个剪刀，他知道怎么拿，也知道应该用怎么样的动作，但是他是从前往后剪，正常用剪刀都是从后往前剪。

这两个例子表明，**即使是非常简单的运动动作，也需要存储知识的指导**。人的大脑对外界信号的理解，比如识别语音、识别文字、识别图片等等，其**识别的最终目的是在我们大脑中提取外界刺激所不包含的信息**。这就是普遍性的知识，只有提取了这种知识，我们才对信号有了理解，对世界有了理解，我们才能做相应的运动动作。

大脑的知识保存在哪个脑区？如下图所示，这是我们看一个词时，大脑的激活状态。大脑活动一开始直接从视觉皮层激活，但其实这个过程不仅仅包含视觉信号加工，大脑活动会迅速扩散到全脑。

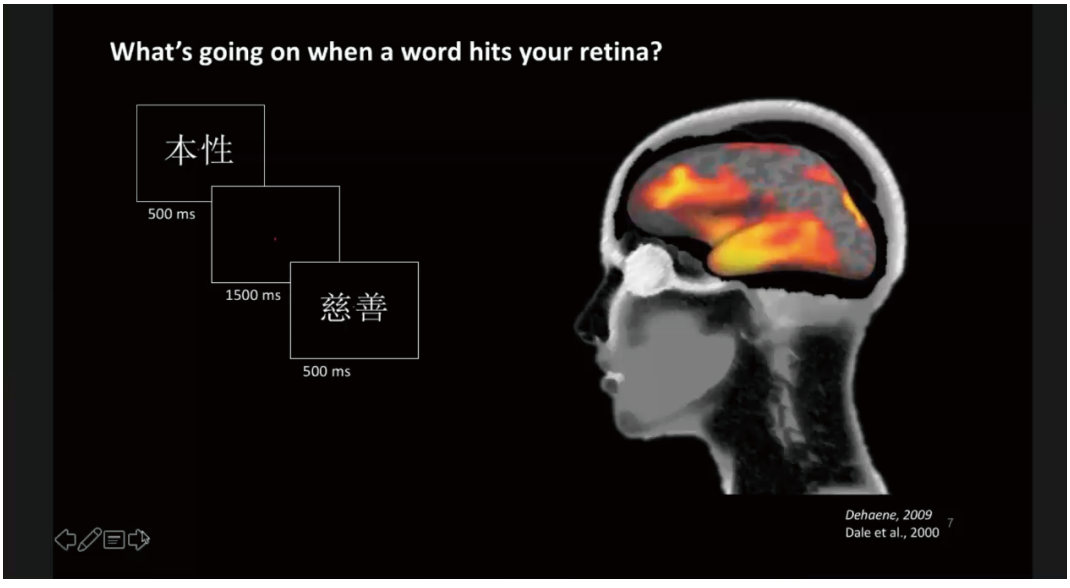


图 2：大脑的激活状态

大脑是一个功能分区非常明显的系统，比如大脑的枕叶处理视觉信号，颞叶处理听觉信号。那么知识存储在哪里呢？综合近二三十年的研究，答案是“**EveryWhere**”。

下图是综合 2009 年之前几百篇研究得到的元分析结果，每一个黄点都是激活点。这是人脑在理解词汇、图片的时候激活的地方，实际上几乎全脑都会被激活，**表面知识可能是非常广泛的分布式存储**。

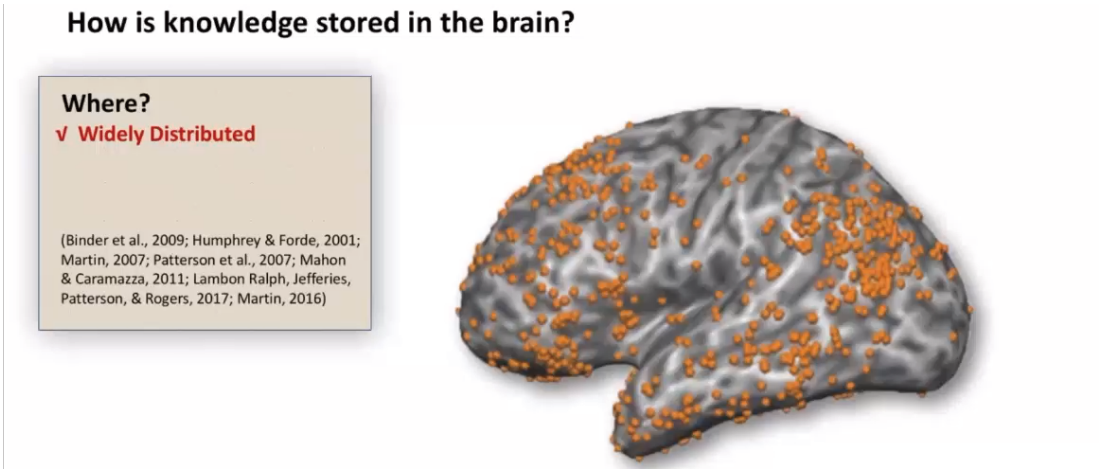


图 3：大脑储存知识时的激活点

大脑的特定脑区保存了什么信息？知识的保存为什么需要这么多脑区的参与？认知神经科学多年来的主流观点是，即使对一个非常简单的概念，比如牛，也分成不同类型的知识存储在相应不同的大脑系统里。

比如听到“牛”这个词，我们会知道它的外形、动作、声音、与人的关系，**不同的信息以相应的感觉经验的模式编码在系统中**。

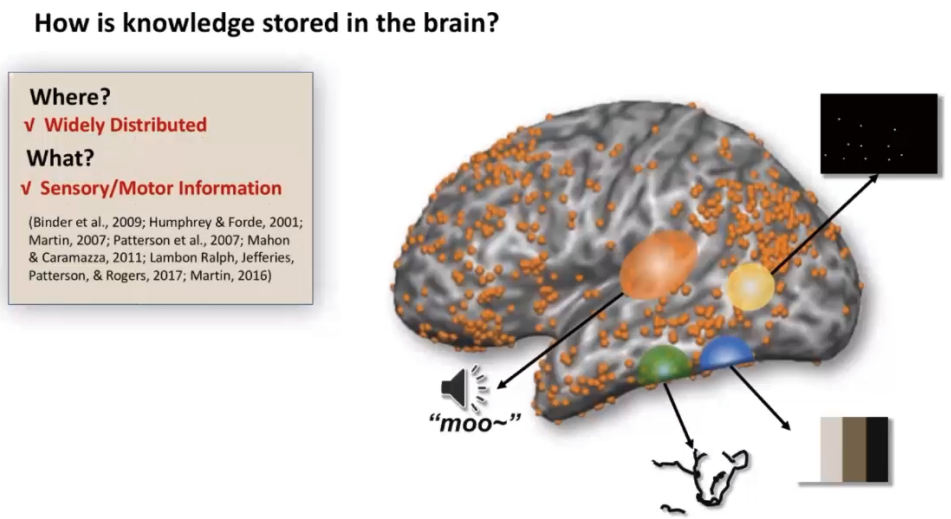


图 4：大脑听到特定词语时的信息处理

其中比如牛的外形，我们的视觉系统看过很多牛，那么相应的激活模式就保存在大脑视觉皮层。下次问我牛的外形，过去的对视觉信号激活的痕迹就会被提取出来，包含它的外形信息。

所以，知识分布式存储的原因是：**第一，简单概念中也包含不同类型的知识；第二，特定类型的知识存储依赖于特定脑区本来的功能。**

二、人类大脑的两种知识表征模式

与 AI 不同，这种人类大脑的知识表征理论中几乎看不到语言的痕迹。我们理解物体、理解语言时，所提取的知识是以对视觉、听觉等信号的感知经验以及与跟对象交互的动作经验信息编码。

那么，这种感觉、运动经验的编码是人类知识表征的全部吗？人又如何存储跟感觉、运动信号并不完全对应的各种抽象知识呢？比如刚才所说的牛，牛肉很有营养、牛会产牛奶等等，这种抽象知识怎么保存在大脑系统里？

我们用实验来回答这个问题。我们通过实验探究先天盲人和正常人在颜色知识表征上的区别。我们大脑中怎么存储玫瑰花是红色这个知识？现有的理论是以过去看玫瑰花的时候，相关的“红色”神经元的发放模式就会印记为“玫瑰”的知识，也就是说印记在视觉皮层里加工形状的视觉编码。

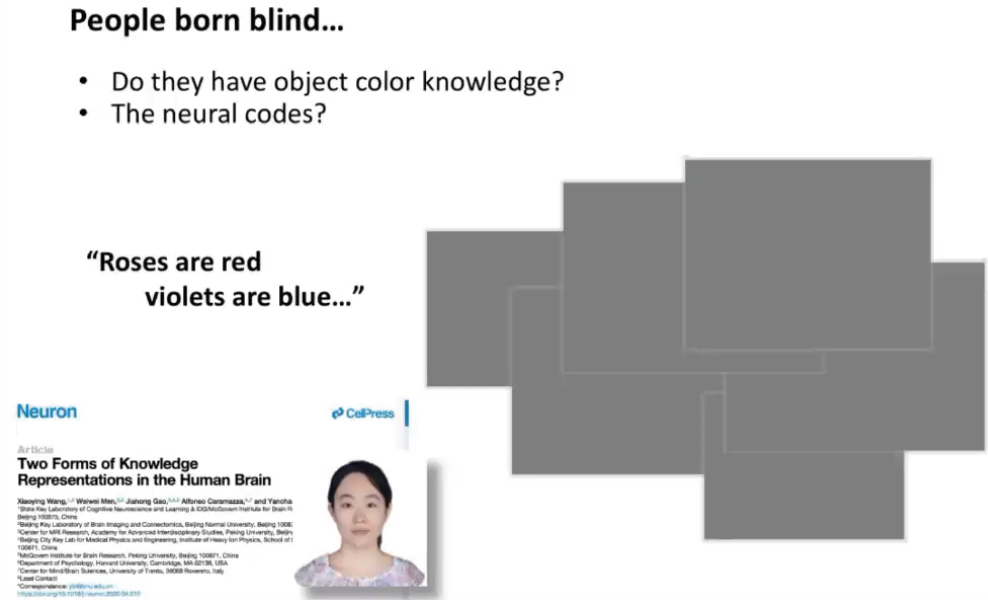


图 5：盲人如何获得不同颜色之间的复杂关系？

先天盲人出生时由于种种原因而没有视觉，问他们玫瑰是什么颜色时，他们仍然正确地回答玫瑰是红色的。颜色是个很特别的特征，因为是光波长度特征，除了视觉没有其他感觉通道可以感知。先天的盲人既然没有视觉经验，只能是靠语言输入获得这个知识。那么他们能获得不同颜色之间的复杂关系吗。

我们首先做了一系列的行为实验。比如，直接问他们不同的东西在颜色上是相似还是不相似、有多么相似。下图是他们行为结果的矩阵图，每一小格都是人们对两个客观颜色相似程度的回答，左边是正常控制组，右边是先天盲人。可以看到**先天盲人不仅仅可以回答颜色知识问题，而且回答的结果模式跟正常人是非常相似的**，相关系数是 0.88。

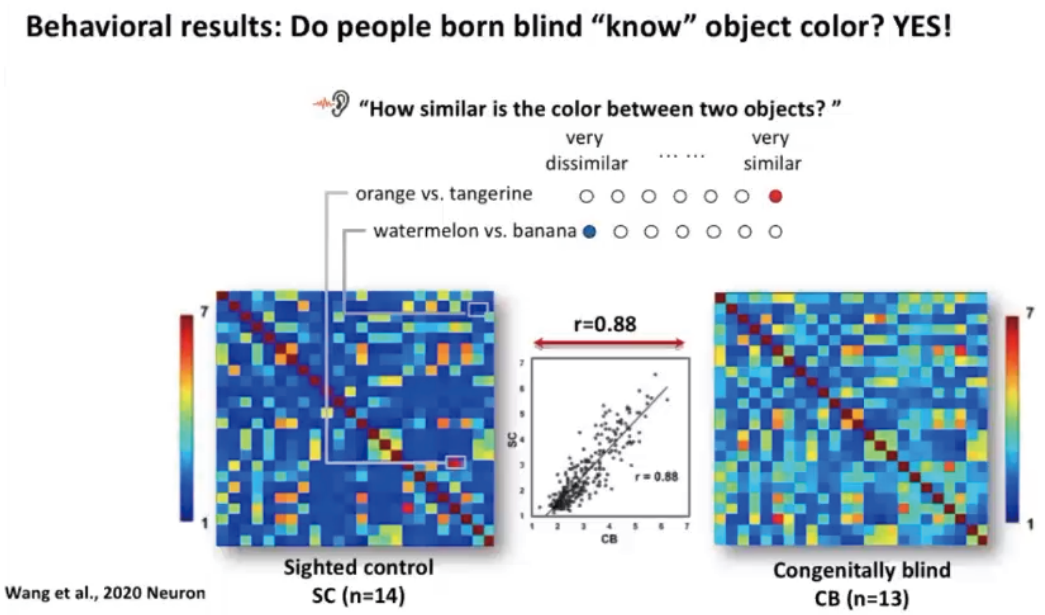


图 6：先天盲人颜色感知的行为实验结果

此外，对于过渡地带的颜色，比如酒红、玫瑰红这些颜色，盲人和正常人的表现也有很大相似性。

下图是物体颜色的判断空间的视觉呈现，左边是正常人控制组，右边是盲人组，可以看到盲人判断颜色之间的远近非常接近正常人。所以即使完全没有感觉到视觉信号，**只提供语言符号信号的话，人也可以建立起相似的知识空间**。

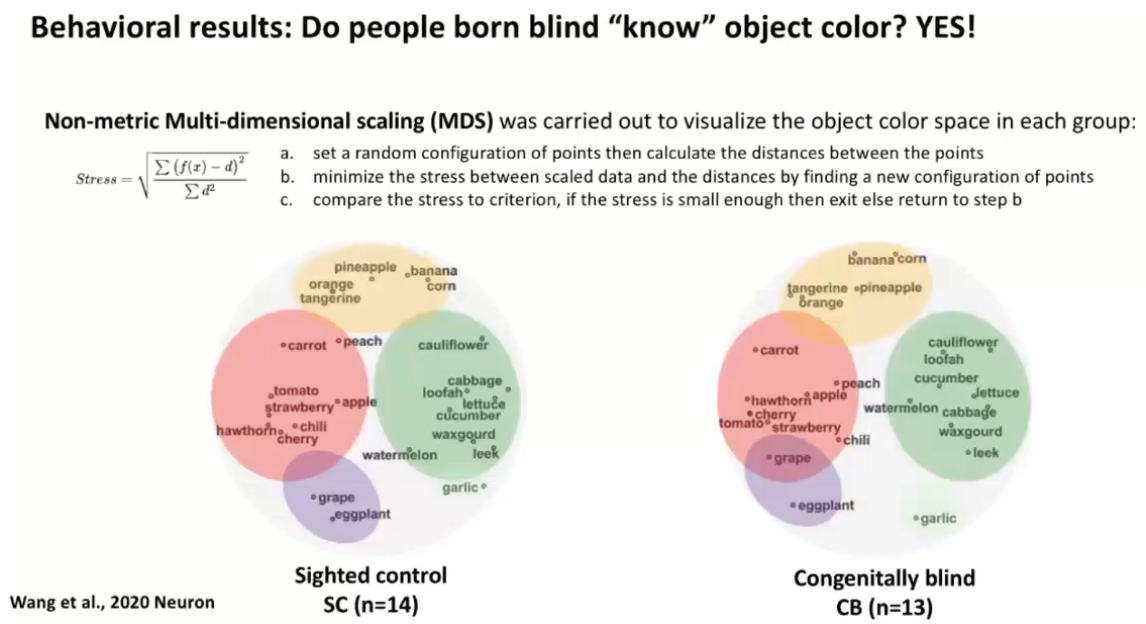


图 7：物体颜色的判断空间的视觉呈现实验

第二个实验探究这两种不同人群在大脑中编码颜色知识的模式。我们把所有的被试放在磁共振机器里，让他们做相似的任务。比如听到苹果、玉米、香蕉这些词，然后回答颜色相关的问题，从而获得每个被试对每个词在回答颜色问题时，大脑所有脑区的激活模式。

通过解码先天盲人和明眼被试人的脑活动对颜色信息编码，发现：第一，我们的确发现**大脑当中有一片视觉脑区负责正常人编码颜色，但是盲人并没有**。该脑区包含只对颜色敏感的神经元，正常人的活动模式是两个东西颜色越像，神经元的活动越像。盲人则没有这个效应，因为他们从来没有颜色视觉经验。

对于盲人而言，在另一个脑区，颞叶前部上侧，神经元的活动模式是两个颜色越像，它们的活动越像。最重要的发现是，不光是盲人，正常人在这个脑区也有一模一样的效应，也就是说**正常人的颜色编码其实涉及两个脑区，一个脑区只有正常人有，以颜色感知觉模式编码颜色知识，另一个区域正常人和先天盲人都有，编码以语言渠道获得的知识**。下图是这种双重编码的知识系统的示意图。**我们把后面这个视觉信号相关的物体颜色知识一个区域叫“ Sensory Derived Knowledge Representation”，前面这个区域 dorsal ATL 叫“ Language Derived knowledge Representation”**。

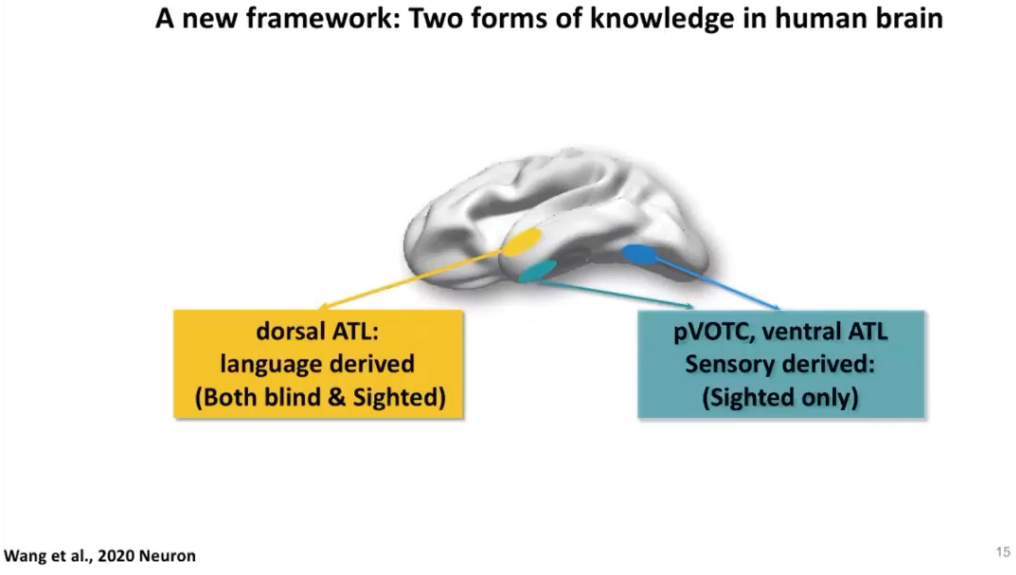


图 8：人脑中知识的两种形式

既然存储有两套不同的编码系统，在大脑不同的区域编码不同的信号信息。大家可以猜测一下，先天盲人怎么表征“彩虹”和“雨”？盲人什么都看不见，但雨还是能感受到的，比如湿度、触觉等等，但是看不见彩虹。**我们这篇在 Nature Communication 2018 年发表的工作发现，对于正常人来说，雨和彩虹非常相似。对于盲人来说，雨是一个具体词，彩虹则是一个非常抽象的词汇，更强存储于完全进行符号编码的脑区，而雨对于盲人而言，还跟正常人一样，在感觉皮层很多区域都有加工**。

我们还可以从另外分布式网络结构的角度问人脑知识表征的问题。再次看看下图，人在理解词汇和图片的时候激活的脑区是分布式的。这个网络有什么结构?

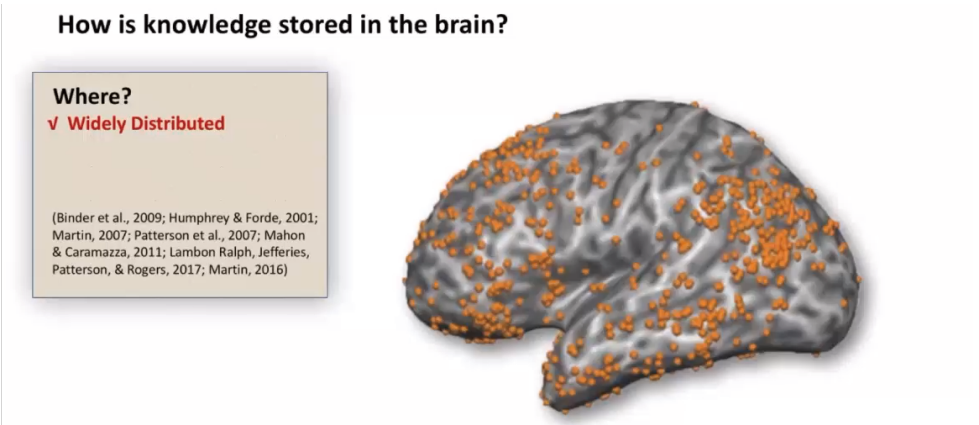


图 9：大脑储存知识时的激活点

我们可以让在被试不做任何具体任务、躺着发呆，然后我们通过功能磁共振测量大脑活动。这时候的大脑活动其实也不是噪音，而是有很多内在规律。我们把不同脑区之间的连接强度提取出来，构成一个由点和边组成的图，就得到了大脑不同脑区之间的连接方式。

我们观察这个图内在的结构，会发现有非常稳定的三个模块—子网络。一个是绿色的脑区之间联系特别紧密，一个是蓝色的脑区之间联系特别紧密，一个是红色的脑区之间联系特别紧密。我们根据以前对这些脑区的理解，发现**绿色的脑区是语言进行加工的地方。红色的脑区是感觉、运动的信号进行加工及多感觉通道融合的地方。蓝色的脑区是执行控制的系统，是对不同的信息进行组合和切换的系统。**

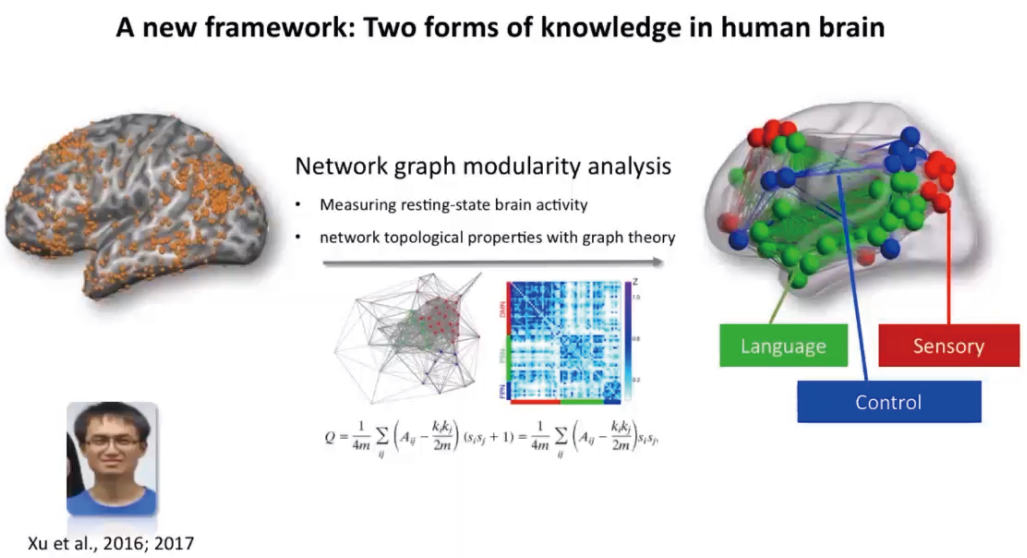


图 10：不同脑区的结构连接

前面发现的提取颜色知识时的两种编码脑区正好就分布在绿色子网络和红色子网络里。所以总体来说，这两个系统在全脑水平上呈现出两个不同类型的网络模块，分别进行语言符号获得知识的编码和感觉获得知识的编码。

我们下面的问题就是，这两个系统的编码机制是什么。我针对每一个系统，举一个实验来介绍一下我们的思路。

三、感觉来源的知识：感觉信号还是计算内容?

首先是感觉来源知识系统，是如何存储知识的？是完全基于感觉信号来编码吗？我们比较了先天盲人和正常人在腹侧视觉皮层上对于形状的加工机制。下图是经典的腹侧视觉皮层，它有典型层级化的结构。**早期的视觉皮层对基本视觉信号敏感，高级视觉皮层会有不同的分区，分别对几个不同重要类型的图片比较敏感，比如人脸、场景、工具、身体等，可能与物体形状知识存储相关。**

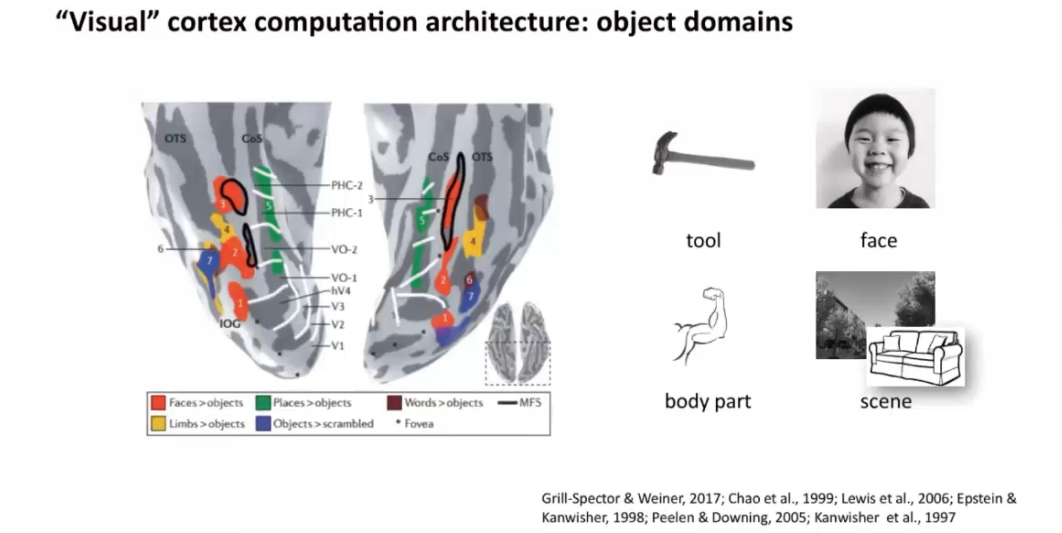


图 11：经典的腹侧视觉皮层

我们比较先天视觉剥夺对这种分布的影响。我们让正常人和先天盲人听很多不同类型的词汇，看看他们视觉皮层激活的情况。**在某一个视皮层区域，正常人在看沙发、办公室等大场景的物体，激活就会特别强。在另一个区域，正常人在看小的工具，比如刚才说的剪子、锤子等，激活就会特别强。对这两个区域，先天盲人的激活模式和正常人是完全一样的。盲人从来没有看见过场景和工具，只能用触觉或者其他渠道获取相关信息，其激活模式也和正常人一样。**

这是不是因为其实光的信号本身并没有那么重要，只要编码相关形状的信息，无论是光信号获得的，还是触觉信号获得的，只要是相似的几何形状关系计算就可以?

此外，**正常人大脑视觉皮层还有一个区域，对动物类的视觉刺激很敏感，比如人的面孔、小猫的形状、小狗的形状，但如果听词和先天盲人听词就没有这种表现。也就是说，大脑的激活模式不仅仅依据对视觉信号的敏感度，还跟物体的类型有关系。**

为什么会有的视觉皮层区域不受感觉信号通道的影响、有些则受？我们推测，这可能与人视觉加工的计算目的相关。生物大脑识别物体的机制不仅仅是为了贴标签。人贴标签是为了交流信息，但是在语言产生之前，人的大脑已经进化了很漫长的时间。在一个简单的场景中，比如餐厅，我们看到的丰富视觉信息中不同元素需要会引导我们作出非常不同反应。看到人要有社会性反应；看到刀叉要有操作性反应；看到桌椅要有绕开或坐下的反应。

视觉系统处理视觉信号，重要目的是正确的提取相关的反应，以适应生存。

人的视觉识别或者视觉知识的存储，会额外考虑到人对应的运动动作是什么。比如下图中的蓝框是人的视觉系统，有不同的层级，这些层级组织的方式要匹配到合适的反应上。

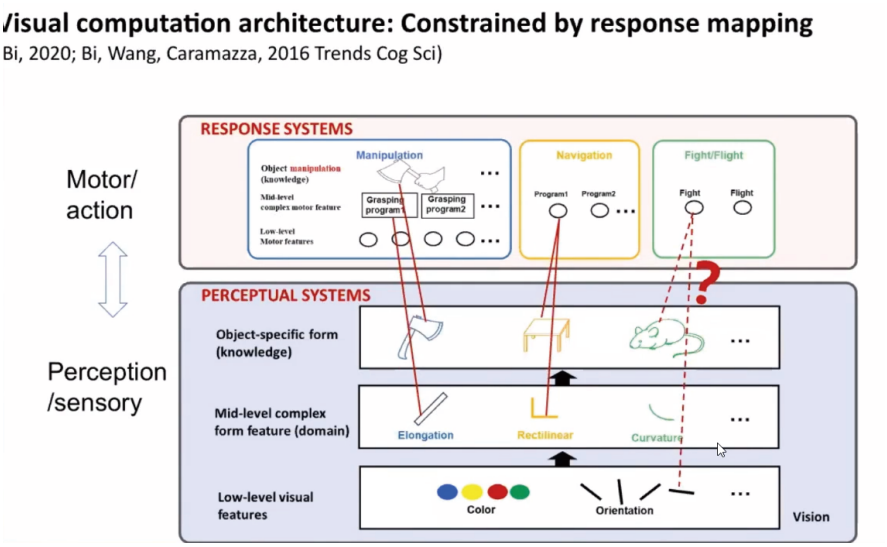


图 12：人的视觉系统层级示意图

在视觉系统的组织或者视觉知识存储的时候，跟反应之间的对应关系就有可能发生在不同的层面。所以，我们可以理解，**对于沙发、锤子等物体，盲人和正常人的视觉组织方式是很相似的，这是因为他们有可能在视觉和运动信号对应上是比较透明的。盲人虽然没有视觉，但是以同样的方式使用这些物体。但是有可能对于蛇、蝴蝶、老虎等等这类信号，并不是从形状上判断如何反应，正常人和盲人接收这些信号的通道不一样，所以正常人和盲人的感觉组织方式就不一样。**

按照特定感觉信息所编码的知识体系，不仅仅是感觉信号本身，还要考虑到不同系统之间的对应关系。所以，**人脑的感觉知识编码和仅对标签分类进行训练的深度学习是非常不同的。**

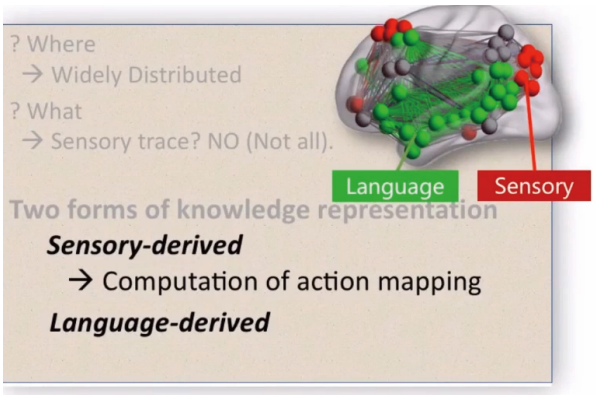


图 13：知识表征的两种方式

四、语言来源的知识和 AI 的知识图谱

先天盲人对物体颜色只利用了语言来构建知识。这是什么编码方式呢？可能与经典知识图谱的问题相似。比如苹果 siri 可以回答语言问题，但是并不直接编码视觉信号、触觉信号等等。

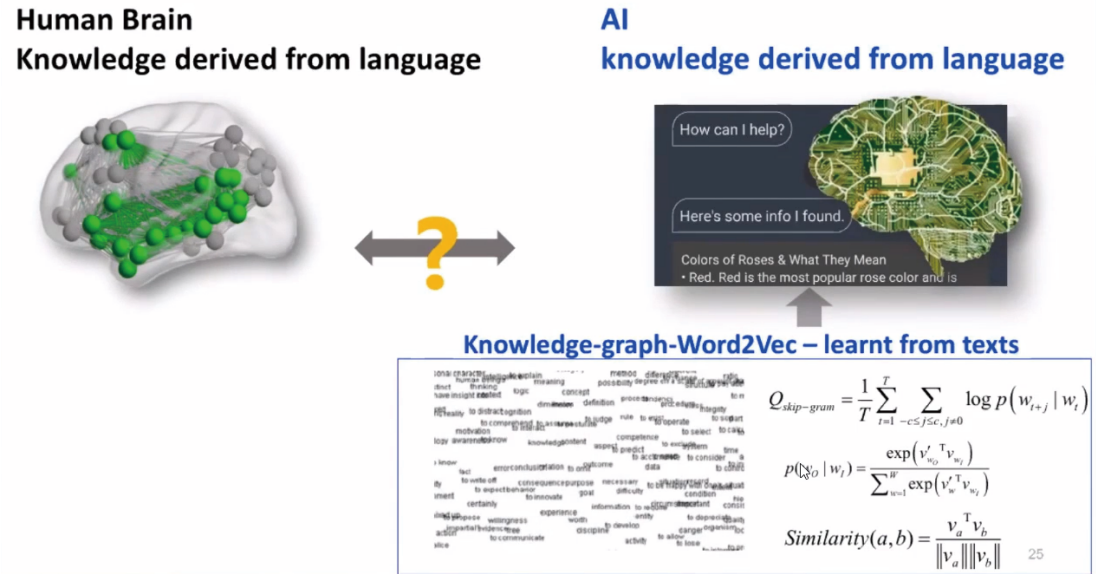


图 14：人脑和 AI 在语言与知识联系上的不同

所以我们做了一个初步的尝试，看它们之间是不是真的有对应关系。我们同时做了两个实验，给定同样的词，一套提供给人的大脑系统，一套提供给 AI 系统，然后建构相关性表征。下图左边矩阵是人在扫描仪里面看那些词（比如赞成、多数、幻想）然后对每一个词获得一个大脑活动的模式。下图右边是我们用 Word2vec 的方式建立的词和词之间的关系。

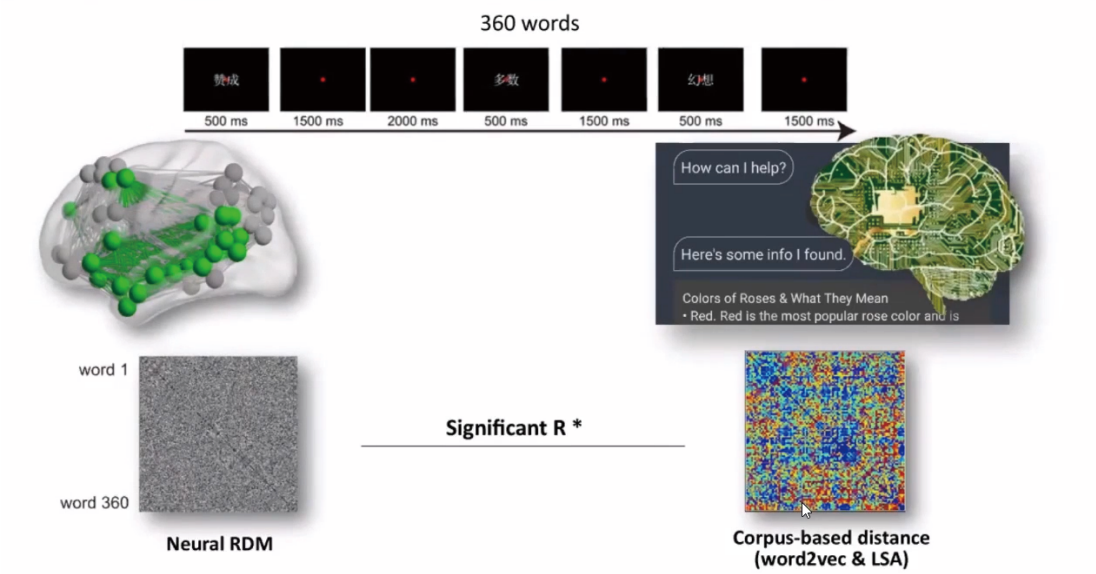


图 15：人脑和 AI 对词语的反应

我们发现人脑几个语言子网络中和 Word2vec 模式唯一有显著相关的就是绿色的系统，也就是人对语言符号加工比较敏感的系统。

大脑系统里可能有这样的绿色系统，它并不关心特定的感觉信号来源比如视觉、听觉、运动等等，但特别对于抽象符号类型的关系很敏感。第一个相关证据是，先天盲人完全没有视觉经验，没有任何其他感觉信号可以获得颜色知识，其编码区域就是在这个绿色系统。第二个相关证据是，绿色系统的活动模式跟 Word2vec 相关，而其它的区域跟 Word2vec 都不相关。

五、总结

我介绍了关于先天盲人的颜色知识、先天盲人的物体形状知识，还有词的计算关系的实验，结论是人脑有两套知识表征模式。一套来源于人特定的感觉神经信号，一套来源于比较脱离感觉经验的抽象语言符号系统。

知识在在人的大脑里以这两套模式存储，组合在一起是人类知识表征。无论我们是看一个图片，还是看一个词，最终都是这两套系统一起激活。需要额外强调的是感觉知识的表征，不仅仅和感觉信号本身有关，还和运动动作相关。我们推测可能感觉来源的知识系统对非文本编码的“Common Sense Knowledge”表征有额外重要的作用；而语言来源的知识系统也在视觉识别中有所影响。

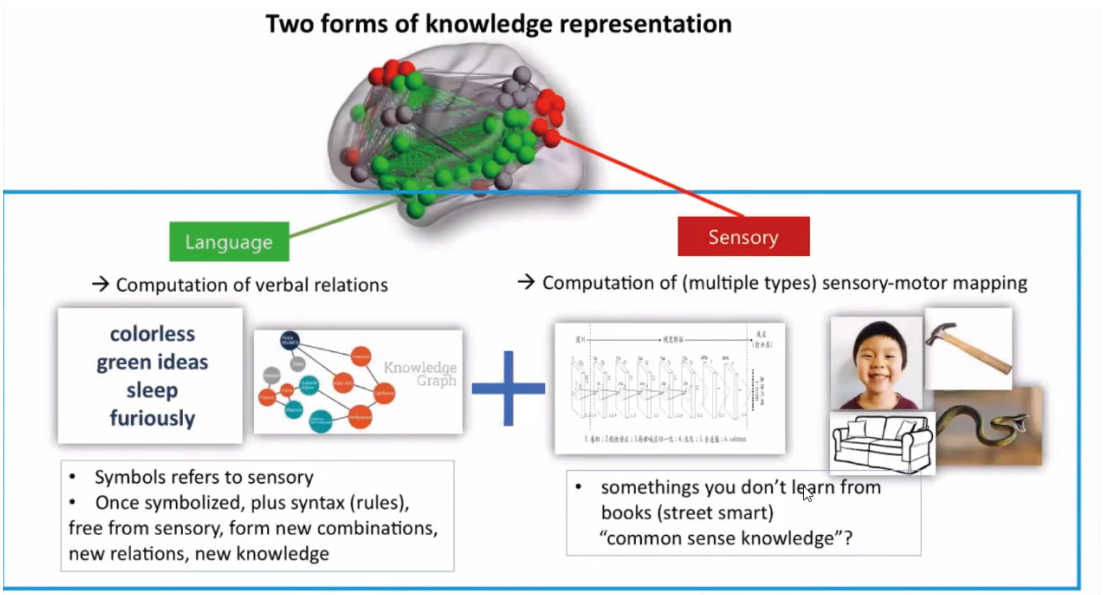


图 16：知识在大脑中的两套存储模式

北大教授方方：人类注意力图和动态机制

转载自：AI 科技评论

6 月 22 日，北京智源大会举行了认知神经基础专题论坛，来自北京师范大学认知神经科学与学习国家重点实验室的毕彦超教授、北京大学心理与认知科学学院的方方教授、清华大学心理学系的刘嘉教授、北京大学计算机系的吴思教授、中国科学院自动化研究所的余山教授分别做了报告，共同探究认知神经科学能为 AI 带来什么启发。

第二位报告者是北京大学心理与认知科学学院院长方方教授，题目为《人类注意力图和功能》。方方在报告中讨论了人脑注意的两个重要属性：**注意力图和动态注意机制。注意力图有两种。注意显著图（Saliency map）源于自下而上的注意，注意优先图（Priority map）则结合了自上而下和自下而上的活动，以及任务相关性。**对多个物体的注意是交替性、节律性、非静态的采样。

以下是演讲全文。

一、注意

我们一般说注意是对外界信息的一种选择性加工。解释注意最好的例子就是交替呈现以下两张图。它们之间有一个非常大的差别，如果不加注意就无法看出。

图 1：它们的差别就在雕像的背后

注意是认知科学里最大的一个领域，每年有超过 1 万篇文章研究注意现象。Corbetta 和 Shulman 在 2002 年描述了关于注意控制的神经模型，总结出两条注意通路。蓝色区域表示背侧额顶网络，负责自上而下的注意控制。 橙色区域表示腹侧额顶网络，负责刺激驱动的注意控制。

Neuroanatomical model of attentional control

(Corbetta and Shulman, 2002)

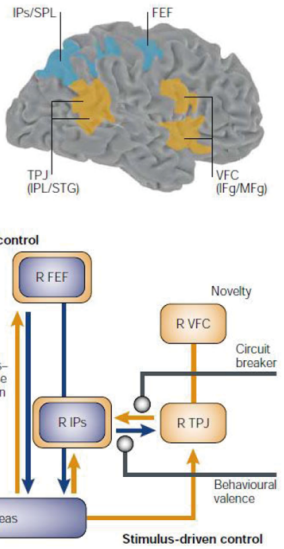


图 2：Neuroanatomical model of attentional control

注意最主要的功能是调节感觉皮层的神经活动，Reynolds 和 Heeger 描述了两种典型方式。**第一种是乘法缩放。对于一个方向选择性神经元，注意可以整体提高神经元在各个方向上的反应。如下图左所示，不注意（蓝线）和注意（红线）之间的变换是一种乘法关系。第二种是锐化。注意可以增强神经元对特定方向的反应，让神经元对外界刺激的选择性更强。这是注意的一些基本功能和神经结构。**

BAAI CONFERENCE 2020北京智源大会

Attentional modulation in visual cortex

(Reynolds and Heeger, 2009)

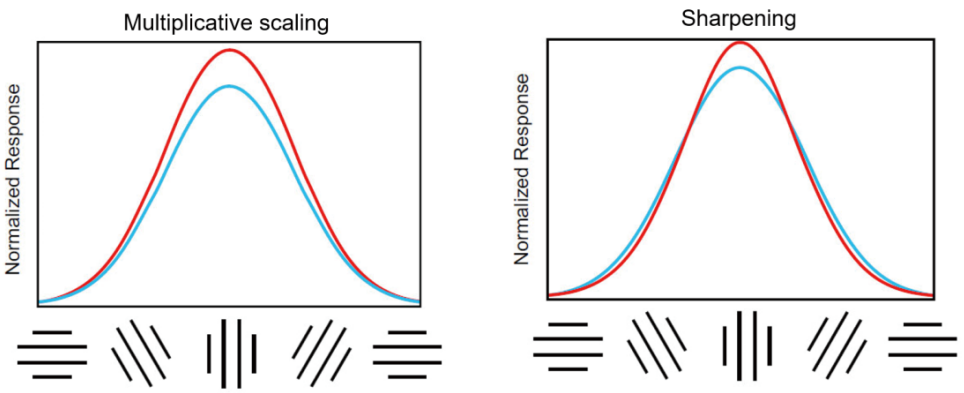


图 3：Attentional modulation in visual cortex

二、注意力图

三维世界投射到我们眼睛上就变成了二维世界，这个二维世界有非常多物体和细节。关于哪些东西更重要的空间分布，就叫注意力图。它分为两种。**一种是注意显著图 (Saliency map)，指自下而上的注意**。例如一个非常奇怪的东西出现在视野中，就会自动吸引你的注意。**另一种叫注意优先图 (Priority map)，则是我们整合自上而下的活动和自下而上的活动形成的注意力图**。做任务时的任务属性也会影响注意放在何处。比如我正在做报告，那么我的注意会更多放在面前的计算机屏幕上。这两种地图如何产生，是我们所关注的问题。

关于 Saliency map ，首先讲最简单的自下而上的注意力图。下图左边是一张海景图，通过计算模型可以算出右边的 Saliency map。越亮的部分表示越有吸引力。

Saliency map

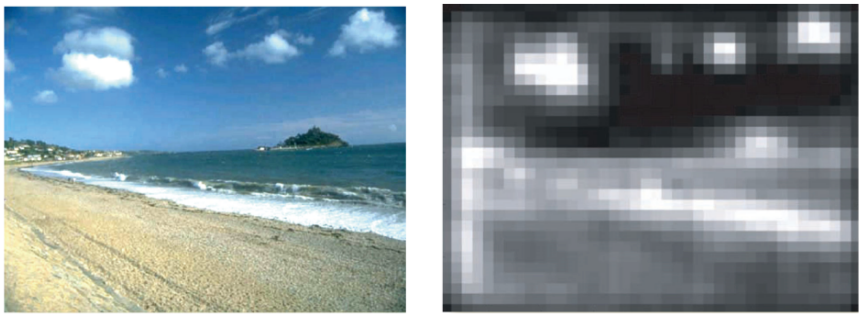


图 3：Saliency map

如何得到右图？我们根据 Itti 和 Koch 在 1998 年提出的模型，计算一张图片在不同尺度上颜色、亮度以及朝向的差异对比度，进行多个尺度的整合，形成 Saliency map。

Itti and Koch's saliency model (IEEE PAMI 1998)

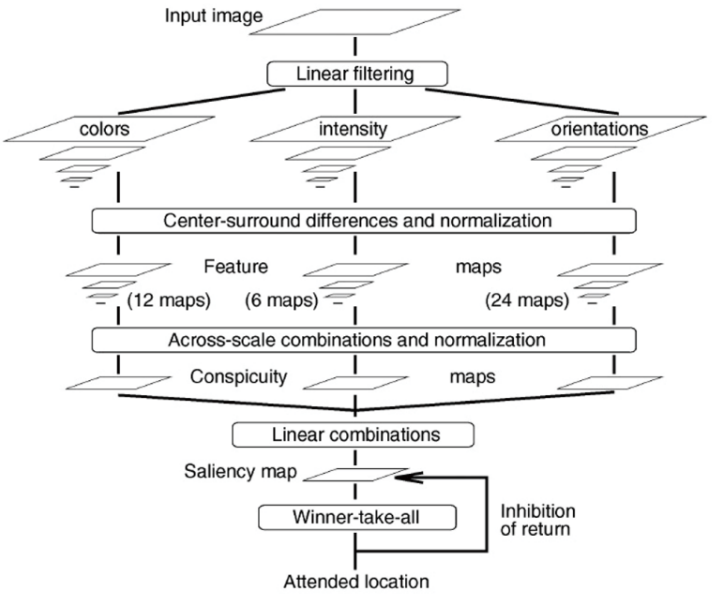


图 4：Itti and Koch's saliency model

有很多重要的文献探讨 Saliency map 在大脑什么地方产生。不同的结论包括在顶叶、前额叶眼区、上丘整合等等。但是我认为视皮层 V1 区就可以充分解释 Saliency map。

为什么以前很多文章都说注意在比较高级的顶叶、额叶等产生？一个可能的原因就是，以前的生命科学研究混淆了自上而下和自下而上的信号。如果我们要研究 Saliency map，必须研究纯粹的自下而上的刺激。怎么样才能做到？我们用无意识的方法，在没有任何自上而下的干扰下，实现研究自下而上的注意。

实验示意图如下。“十字”是参与者的注视点，四个“减号”是 Saliency map 的位置。它会显著吸引我们的注意，而且经过实验操纵后不会被意识到。我们改变“减号”的角度，将“减号”和“I”之间的夹角分别设为 0 度、15 度、30 度和 90 度。随着夹角增加，它吸引注意的能力逐渐增强。

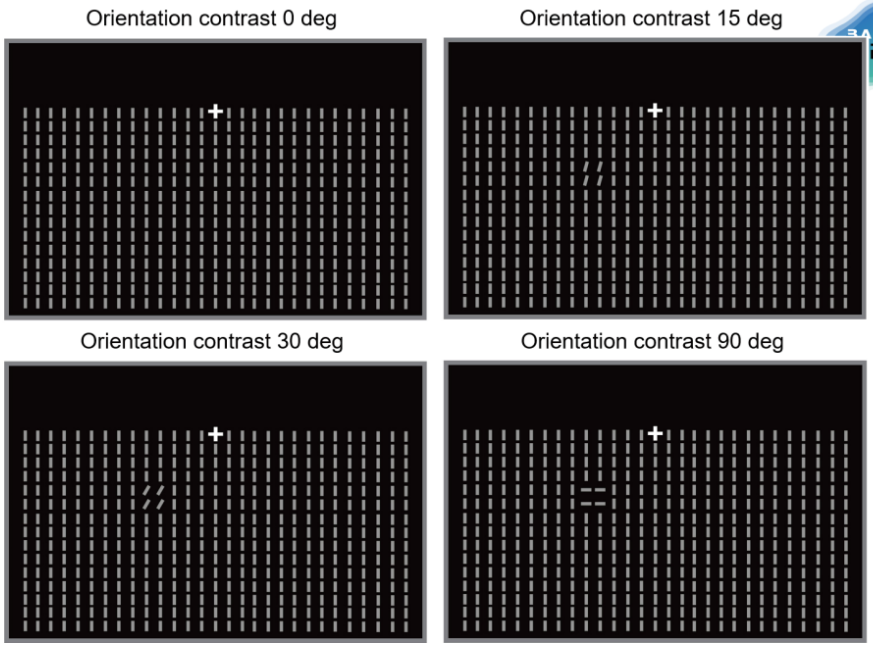


图 5：自下而上的注意研究示意图

如何把这个刺激变得无意识？下图是我们的实验流程。首先呈现线索图片 (Cue) 50ms，然后呈现掩蔽图片 (Mask) 100ms，然后呈现注视点 50ms，最后是探测任务，探测第四张图十字下面两个点的相对位置。由于线索仅仅呈现非常短的 50ms，又紧跟着 100ms 的掩蔽图片，所以被试完全不会意识到线索的存在。但是探测任务**放在线索的显著区，被试依然有较好的表现。任务放在对侧的话，被试的表现就比较差。两个条件的差别就代表自下而上的注意强度**。

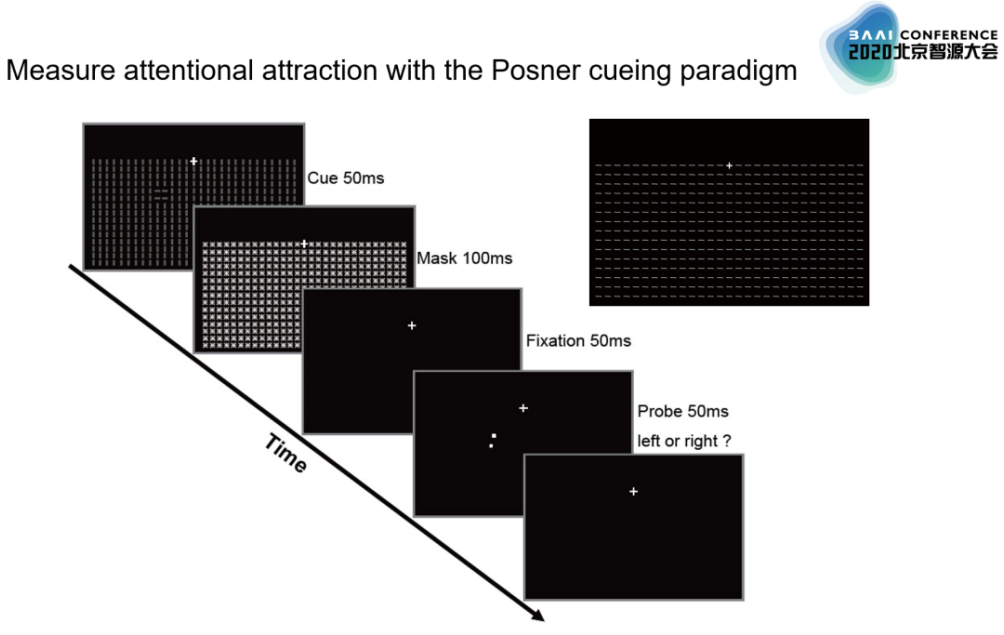


图 6：Measure attentional attraction with the Posner cueing paradigm

我们接着利用视皮层 V1 区神经元的属性构建注意模型。看看下图的数据，随着朝向倾斜角度增加，注意的吸引力也逐渐增强，跟计算模型吻合得非常好。我们发现 **Saliency map 跟 V1 的神经元活动是有关系的**。

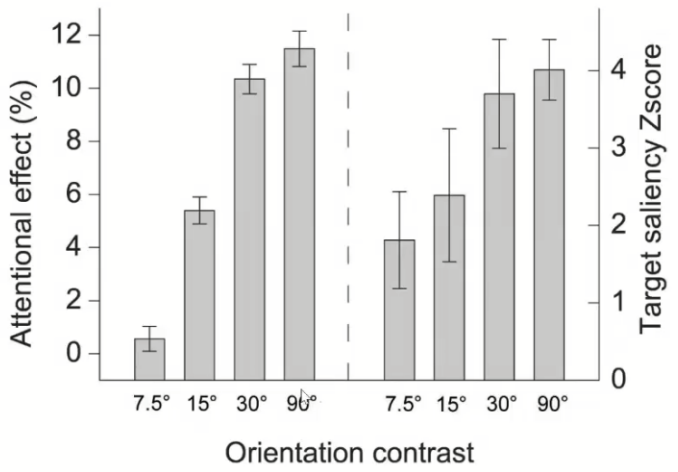


图 7：Saliency ma 与 V1 的神经元活动关系图

从初级视觉层到顶叶，Saliency map 的效果逐渐减小。最明显的是 V1 区域，可以产生自下而上的注意。我们上面的研究基于人工刺激，下图则基于自然场景。图中的马具有非常高的显著度，甲壳虫具有较低的显著度。他们在大脑皮层诱发出的信号有没有区别?

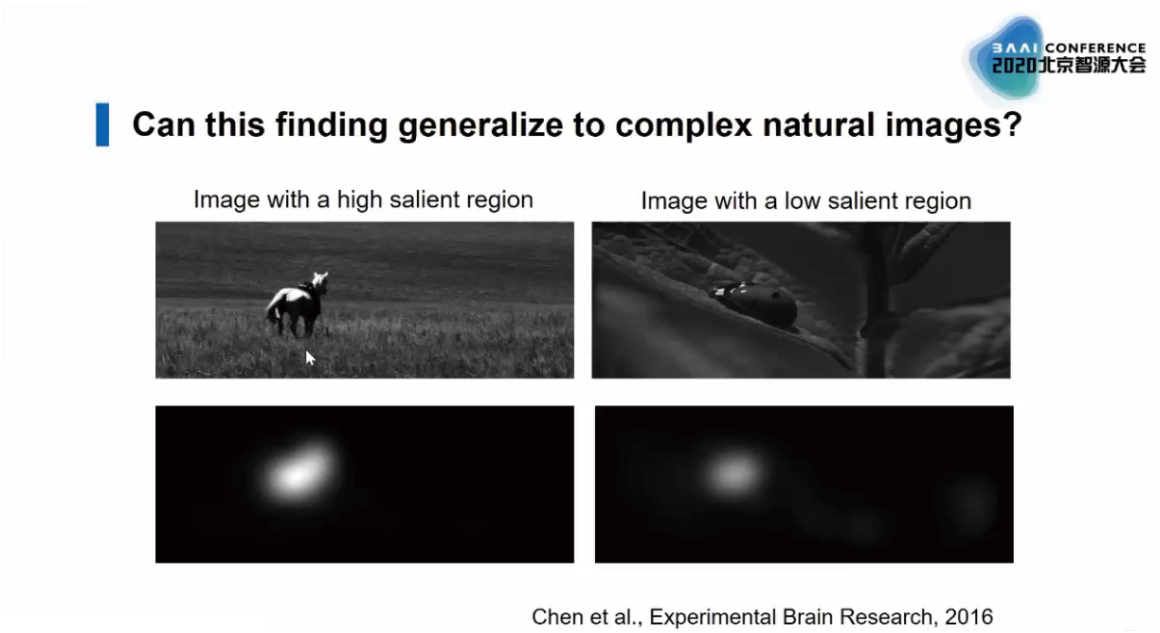


图 8：大脑研究实验

我们重复了这个行为学实验，发现马确实可以诱发出更强的注意信号，甲壳虫则不可以。并且还是在 V1 区域展示了注意的分布，所以我们再一次用自然场景证明了 Saliency map 跟 V1 是相关的。

基于这个生理学依据，我们构建了一个动态注意模型。这个模型的大致框架有三个组件。第一部分参考感觉反应，模拟 V1 神经元对自然场景做稀疏编码。第二部分是中央凹图像多分辨率金字塔方法。对于自然场景，如果盯住这个红色十字注视点，编码会非常清晰，但是对外围的编码就非常粗糙，第三部分模拟视觉工作记忆，注视一个场景后很难立刻再跳回去。

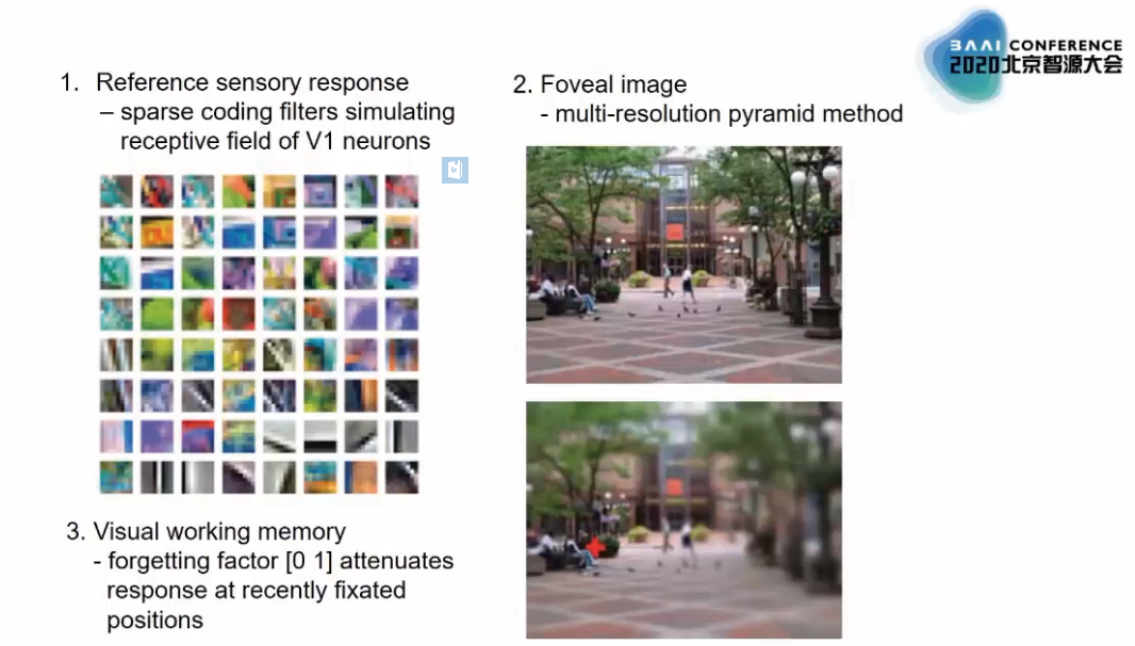


图 9：模型框架的 3 个组件

我们把这三个组件放在注意模型里，构建了一个基于图论的模型。这个动态的注意模型将 V1 神经元构成网络，用该网络搜寻图片上最富有信息的区域，然后跳到第二富有刺激信息的区域。下图中最下方图的红线代表在自然场景里人类的眼球运动轨迹，中间是我们模型预测的轨迹。实验结果表明我们的模型和实际情况吻合得更好。

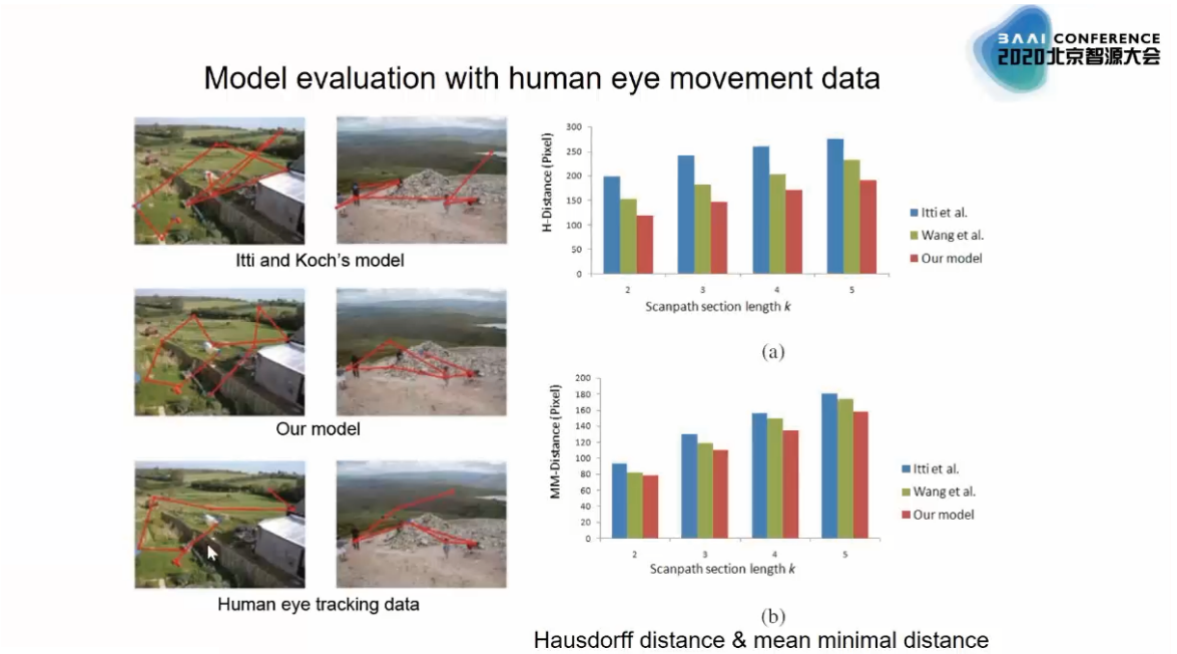


图 10：眼球运动数据的模型评估

关于 Priority map，回到那张海景图，Saliency map 是中间上图。任务要求寻找图上的小岛，于是小岛被高亮标记。中间这两张图并在一起后，小岛应该仍是高亮的。**Priority map 整合了自下而上的显着性，与当前任务的相关性**。

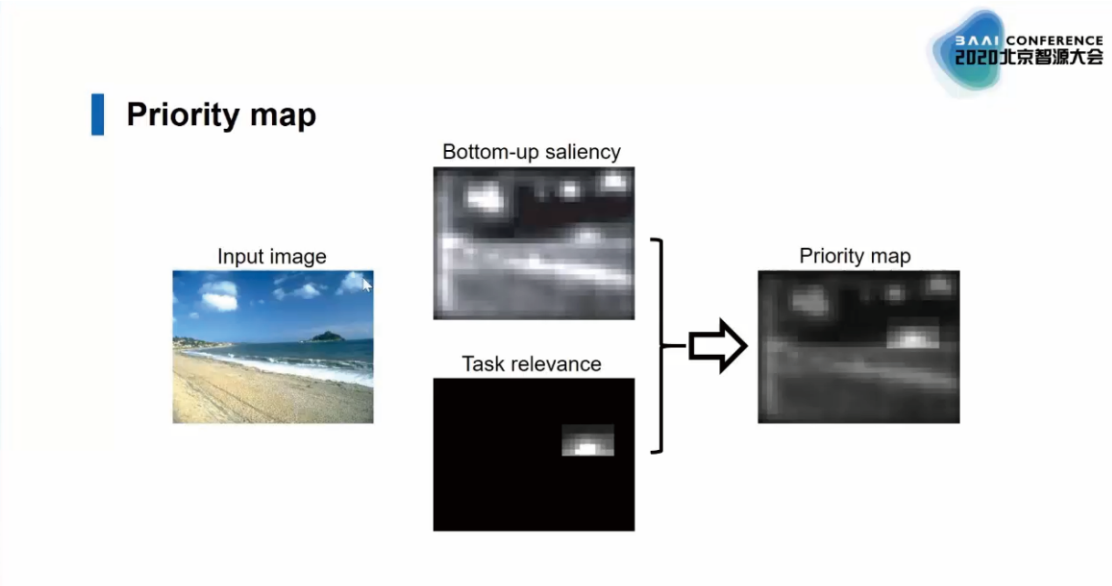


图 11：Priority map

我们又用人的面孔进行实验。面孔比人工刺激复杂得多，还具有倒立效应，即同样的脸倒过来后很难识别。这也非常影响 Priority map 在面孔上的分布。

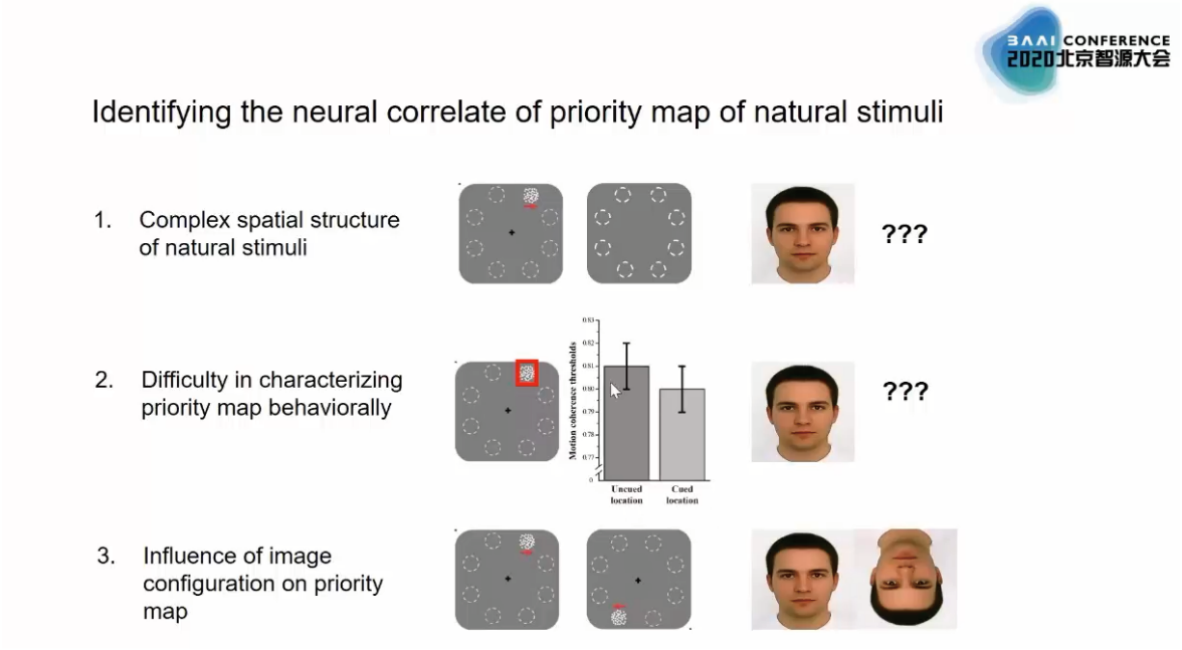


图 12：Priority map 在面孔上的分布

我们给被试看正立脸、倒立脸和相位打乱的面孔，让被试的眼睛在面孔上随便跳动。另外，扫描被试视皮层对面孔的反应，得到行为学的数据和脑活动的数据。

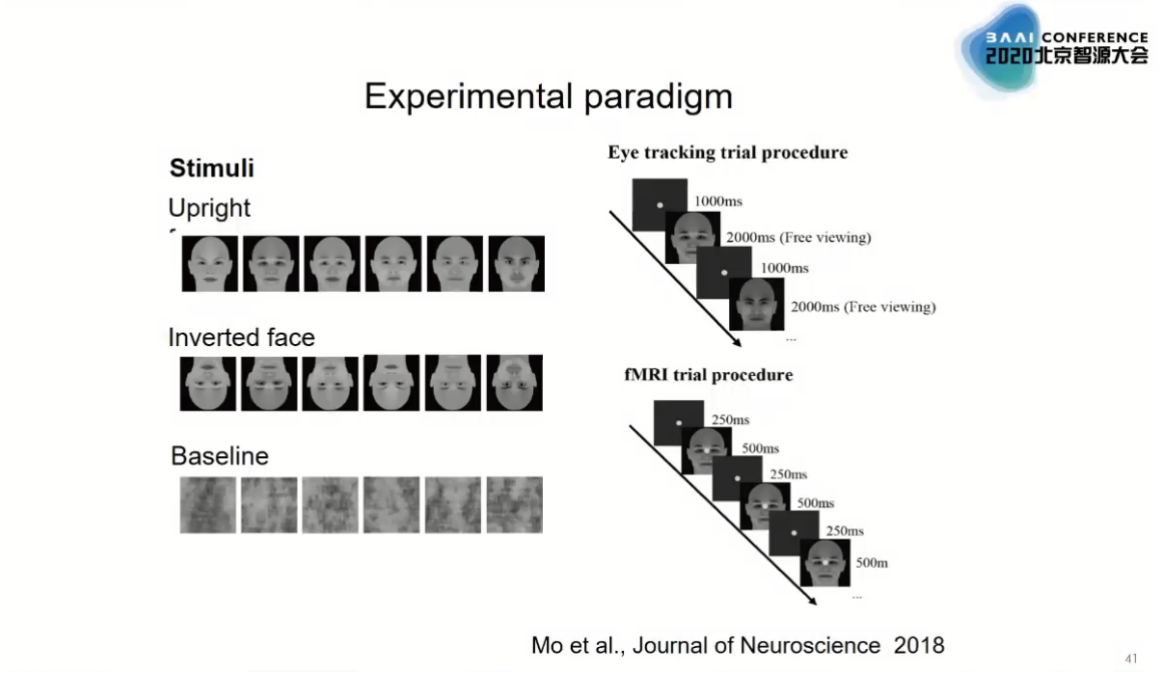


图 13：不同刺激下行为学的数据和脑活动的数据

我们重构出任意一个视皮层对面孔每一个部分的反应。下图右下角是模型重构的反应，颜色越暖说明视皮层相应区域对面孔的反应更强。右上角是行为学数据，我们第一眼看面孔时注视什么地方。颜色越暖说明第一次着眼此处的概率越大，也就是该区域越容易吸引眼球。

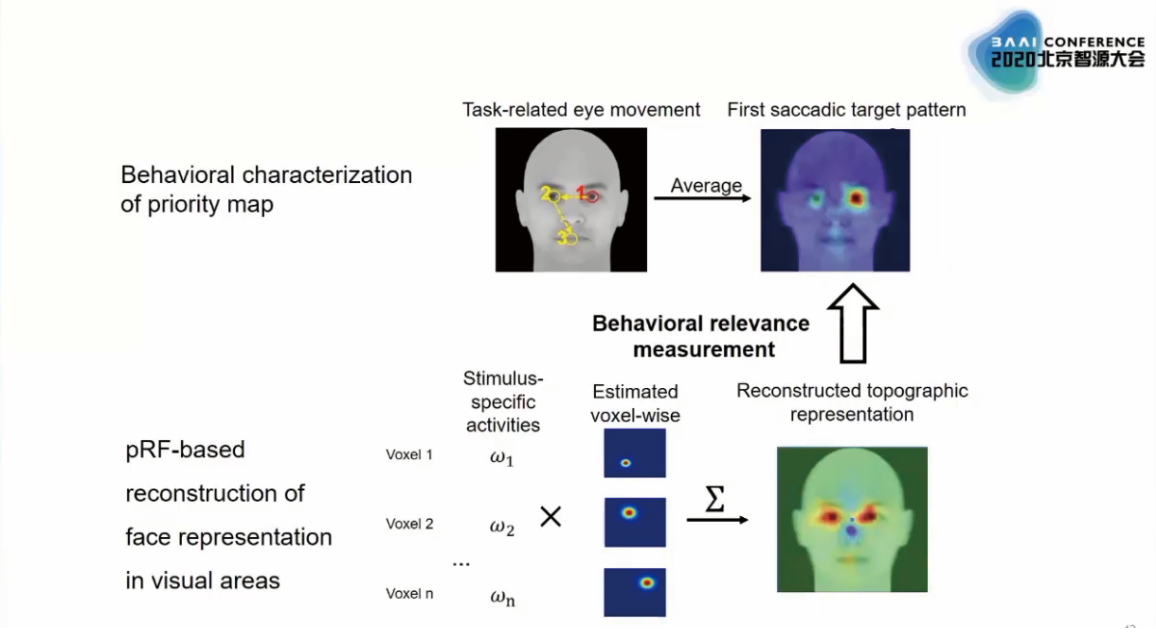


图 14：模型重构的反应

下图是这个实验最主要的结果。我们测量最左边正脸和倒脸吸引眼球的程度，描述了视皮层 V1、V2、V3 区域对正脸和倒脸反应的分布。

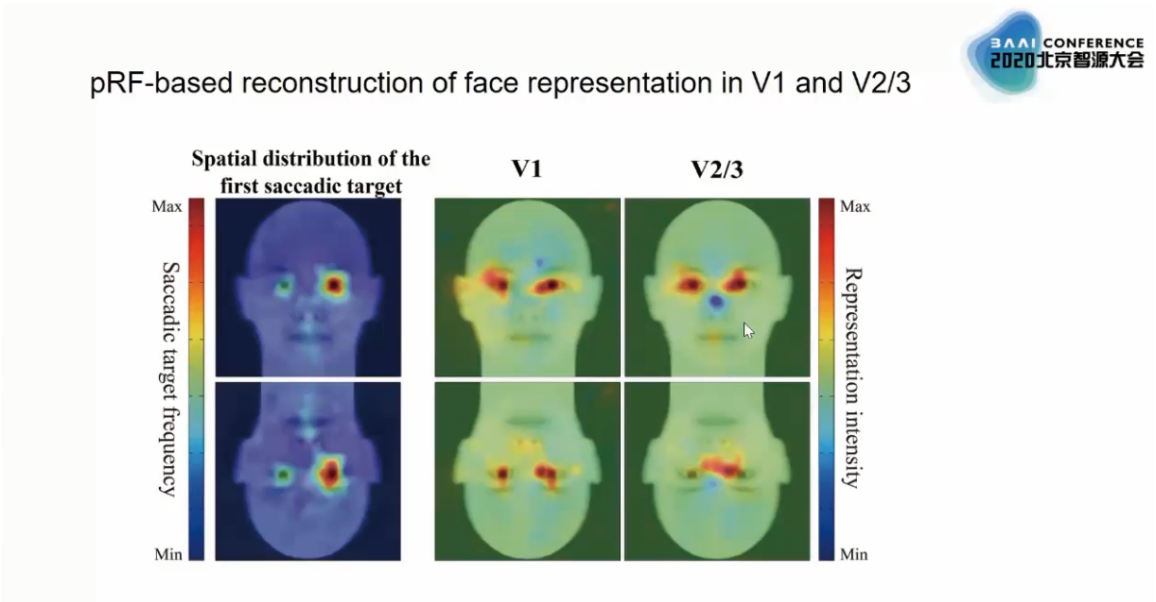


图 15：正脸和倒脸吸引眼球的实验

我们发现 V2 和 V3 对正脸的表征是最精确的，远远高于其他三种情况。V1 对正脸和倒脸的反应表征的精准度都比较低，但是 V2 和 V3 对正脸表征的精准度比对倒脸表征高很多。

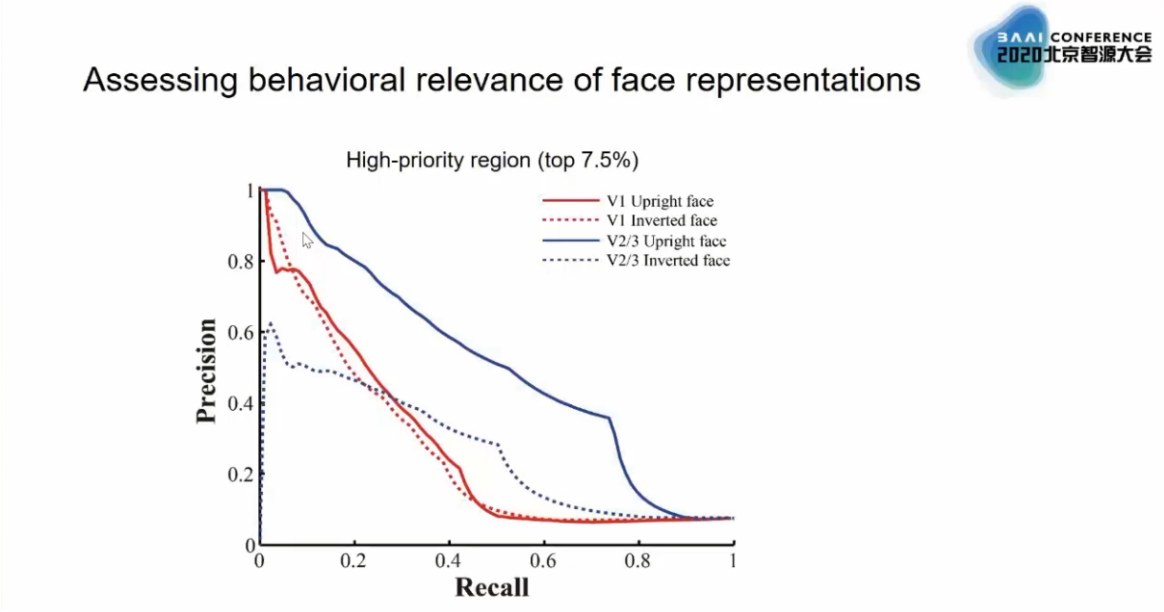


图 16：Assessing behavioral relevance of face representations

总结一下就是，**人类早期视皮层，从 V1 区域到 V3 区域，V1 对 Saliency map 即自下而上的注意起到很好的表征作用，V2 和 V3 则对 Priority map 即自上而下的、任务驱动的注意起到很好的表征作用**。

三、注意的动态机制

过去关于注意采样的大部分观点认为，我们一旦注意到一个物体，对它的注意是持续的、静止的。但事实是不是这样的？从现在的数据来看，不一定。

另外一种观点是有数据支持的，特别是同时注意两个物体的时候。如下图所示，一种理论提出注意把关注点分割为两块，同时关注蓝色和黄色方块，这是一种平行和稳定的关系。另外一种理论认为，注意在这两个物体之间切换。我们希望用实验来提供进一步证据。在我们的脑成像实验之前，行为学研究已经发现，如果同时注意左右两个物体，注意其实是左右切换的、顺序的、周期性的交替采样过程。我们的脑成像结果也证明，**对多个物体的注意是交替性的、节律性的采样，而不是一种静态的过程**。

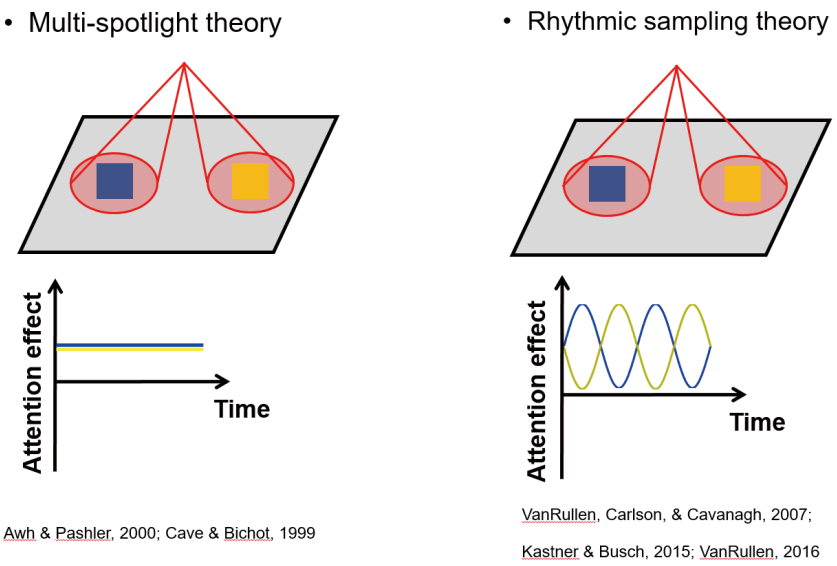


图 17：Mechanism of concurrent multi-target attention

下图表示随着不同的任务要求（100% 注意 A 并且 0% 注意 B、75% 注意 A 并且 25% 注意 B、50% 注意 A 并且 50% 注意 B），注意在不同的物体之间节律性分配。它不仅仅对静态物体有用。对于动态的物体，比如两个运动的小球，同样可以发现类似的节律性采样过程。

BAAI CONFERENCE 2020北京智源大会

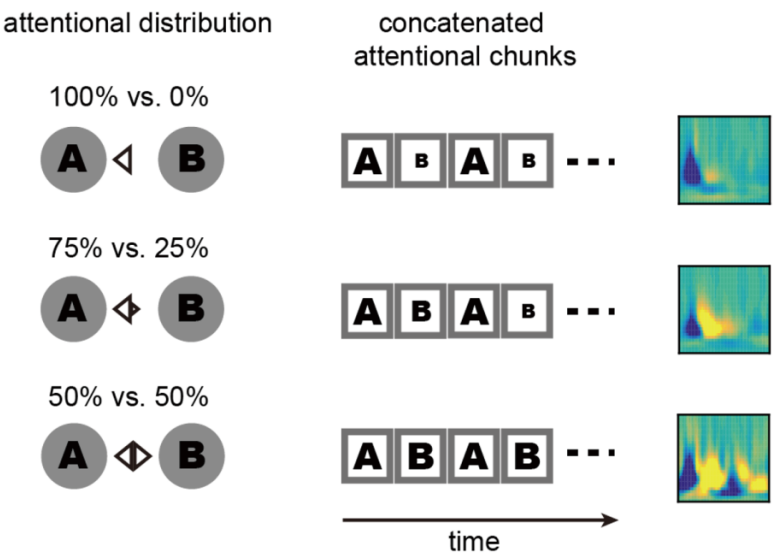


图 18：同时注意多个特征时，采样的具体过程

上面讲的是对于物体的采样和对于空间的采样，如果同时注意多个特征，采样是怎么样的过程呢？这个问题相当复杂。比如对于任何一个朝向、任何一个运动方向来说，有很多神经元同时进行反应，怎样描述这种同时的反应？我们做了一个脑磁实验，呈现一个刺激，测量脑磁信号。这些信号由大脑中不同朝向神经元的不同通道反应组合而成。然后我们用脑磁信号反解出每个通道的反应。

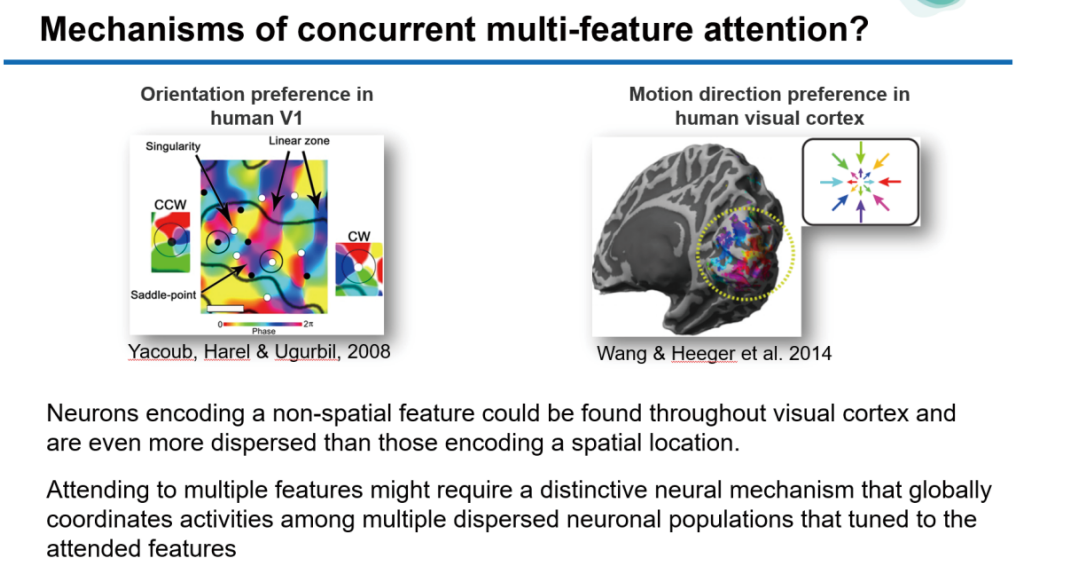


图 19：Mechanisms of concurrent multi-feature attention

实验表明，如果我们同时注意两个特征，对这两个特征的表征同样是交替性的，而且是反相位的。无论是基于空间的注意、基于客体的注意，还是基于特征的注意，都不是静态的过程，而是在不同的空间、客体和特征之间交替。

清华大学教授刘嘉：从认知到计算：认知神经智能科学

转载自：AI 科技评论

6 月 22 日，北京智源大会举行了认知神经基础专题论坛，来自北京师范大学认知神经科学与学习国家重点实验室的毕彦超教授、北京大学心理与认知学院的方方教授、清华大学心理学系的刘嘉教授、北京大学计算机系的吴思教授、中国科学院自动化研究所的余山教授分别做了报告，共同探究认知神经科学能为 AI 带来什么启发。

第三位报告者是清华大学心理学系教授刘嘉，题目为《从认知到计算： 认知神经智能科学》。在报告中，刘嘉教授首先回顾认知科学的历史，解释打开人脑黑箱的意义，然后通过一系列认知神经科学的实验范式和研究技术，揭示了深度神经网络的内部表征与算法以打开 AI 的黑箱，展示了人脑与类脑双脑融合的可能路径。

以下是演讲全文。

今天我的报告主要围绕如何从认知神经科学对大脑的研究方法论，来理解深度神经网络的工作方式。

一、行为主义

在 AI 里，我们通常会遇到图片识别的问题，我们把图片输入到训练好的 CNN 里，CNN 告诉我们这是一匹马。这个过程是我们现在主流的深度神经网络所做的工作，采用**行为目标导向**，即在输入端和输出端建立关联，而把中间过程当成一个黑箱（Black Box）。

显然作为科学家，我们肯定有兴趣把它打开，但是问题是有必要吗？打开和不打开究竟对理解 AI 以及推动 AI 发展有没有帮助？

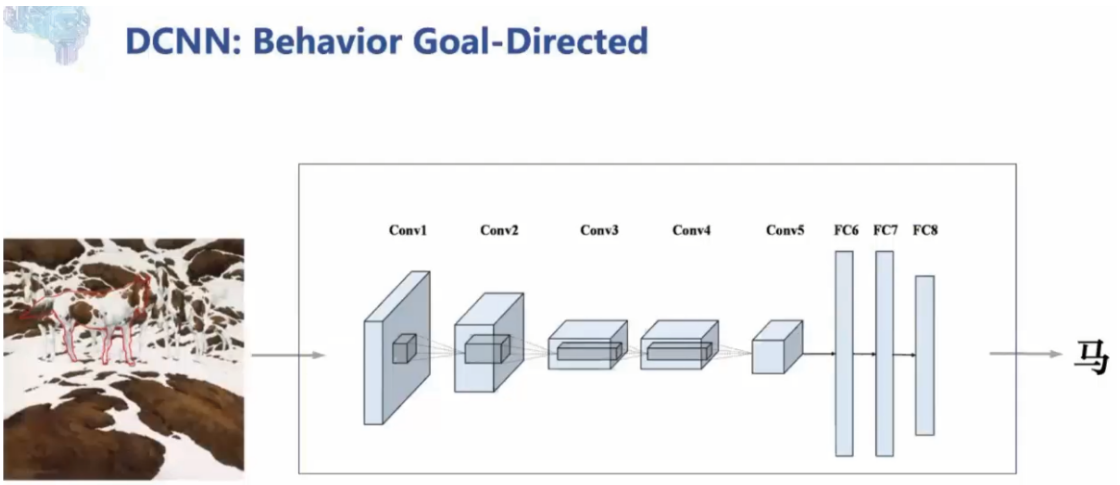


图 1：DCNN：Behavior Goal–Directed

在心理学历史上也曾有类似的争论。关于刺激和行为之间关系的研究最早是由 Pavlov（巴普洛夫）开展的，他称之为条件反射。即当铃铛和食物同时出现或者铃铛比食物稍微早一点出现的时候，这时候就可以建立刺激与行为的联系。即当食物不出现时，仅仅摇一下铃铛，狗也会分泌唾液。至于狗的大脑里面发生了什么，当时大家认为不重要，当成黑箱就好；而我们需要关注的是刺激和行为之间连接的法则。

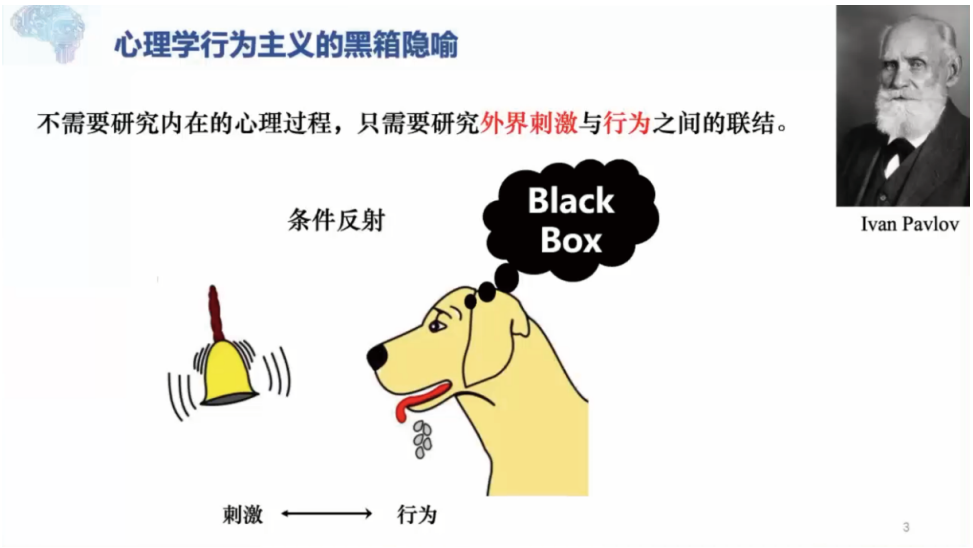


图 2：心理学行为主义的黑箱隐喻

这个观念从上个世纪三十年代一直到五六十年代都占据着心理学的主要地位，称为行为主义。行为主义有一个著名的黑箱隐喻，即行为主义代表人物 Watson（华生）说过：“**给我一打健康的婴儿，一个由我支配的特殊环境，让我在这个环境里养育他们，我可担保，任意选择一个，不论他的父母的才干、倾向、爱好如何，他父母的职业及种族如何，我都可以按照我的意愿把他们训练成任何一个人物——医生、律师、艺术家、大商人，甚至乞丐或强盗。**”

这句话背后的逻辑就是深度神经网络的“行为和目标导向”，翻译成心理学的术语就是“人是环境的产物”或者“智能是环境的产物”。

二、Garcia 效应

但是理解外部环境和行为之间的关系就够了吗？后继的研究表明这远远不够。Garcia（加西亚）曾经研究放疗所产生的负作用，如恶心呕吐等。具体而言，他给老鼠进行放疗，然后观察放疗之后老鼠的行为。Garcia 发现了一个非常奇怪的现象，放疗后的老鼠中有一些老鼠开始拒绝喝水，再渴也不不喝水。Garcia 深入了解后发现，那些拒绝喝水的老鼠的盛水容器是塑料瓶，而继续喝水的老鼠的盛水容器是玻璃瓶。

玻璃和塑料之间有什么区别？非常简单，因为玻璃瓶是没味的，而塑料瓶是有味的，也就是说老鼠把它恶心呕吐的症状和塑料瓶的味道联系在一起了，老鼠会“认为“自己呕吐是塑料瓶带来的。从表面上来看，这就是一个非常简单的刺激（塑料瓶的气味）和行为（呕吐）之间的联结，也就是我们刚才说的条件反射。但是！ Garcia 进一步发现，当他用类似气味的条件，比如闪光、铃声来试图形成老鼠不喝水的条件反射，发现怎么都建不成联结。也就是说老鼠只能把气味和它的呕吐建立联结，而不能把闪光、铃声来与它的呕吐建立联结。基于此，

Garcia 用生物准备性（Biological Preparedness）的概念来对行为主义提出了挑战。

生物准备性的核心有两点：第一，不是所有的刺激都能和反应建立联结；第二，有机体的学习潜能都被其生物学基础所约束。也就是说黑箱里面的东西制约了刺激和反应联结的形成。

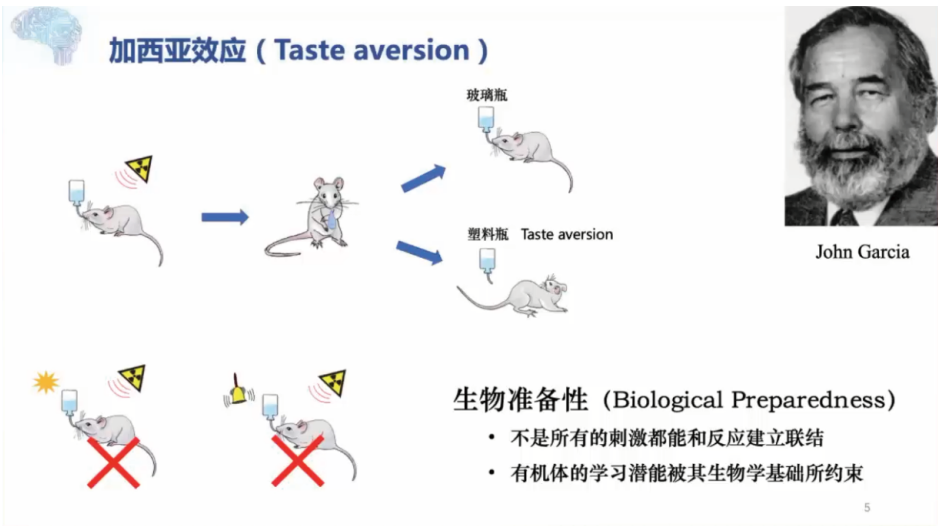


图 3：生物准备性

正是 Garcia 这个实验使得我们开始研究老鼠的大脑里在“想”什么，狗的大脑里在“想”什么，于是认知科学由此诞生。科学家开始逐渐把大脑的黑箱打开，知识表征、注意力等概念就是认知科学在研究大脑机制时提出的认知概念。以前行为主义认为人只是环境的产物，现在我们知道，**人不仅仅是环境的产物，而且也是环境的营造者，人有其自身的内部加工过程。同样，深度神经网络的内部表征与算法也必然影响刺激与行为的连结，也必然决定其智能的形态和本质。**

图 4：认知科学

之后认知科学和神经科学产生连接，我们开始了解认知模块和表征的生物学基础。基于认知神经科学过去20–30 年的工作，我们开始理解视觉的产生机制。首先是初级视觉过程，对物体的线条、颜色、对比以及运动等特征进行初步分析。接下来是中级视觉过程，我们开始把物体从局部的信息整合成形状、表面、深度信息，最后我们把这些信息整合起来进入高级视觉过程，这时候我们就可以实现物体识别等。

认知神经科学帮我们打开了大脑黑箱的一部分。那么我们为什么不用认知神经科学的方法论和工具，来理解人工神经网络的功能模块和内部表征，了解人工智能背后的智能本质，获得可解释、可预测的 AI？这里，我把这个思路称为人工智能的认知神经解析，即用认知神经科学的方法来研究 AI。

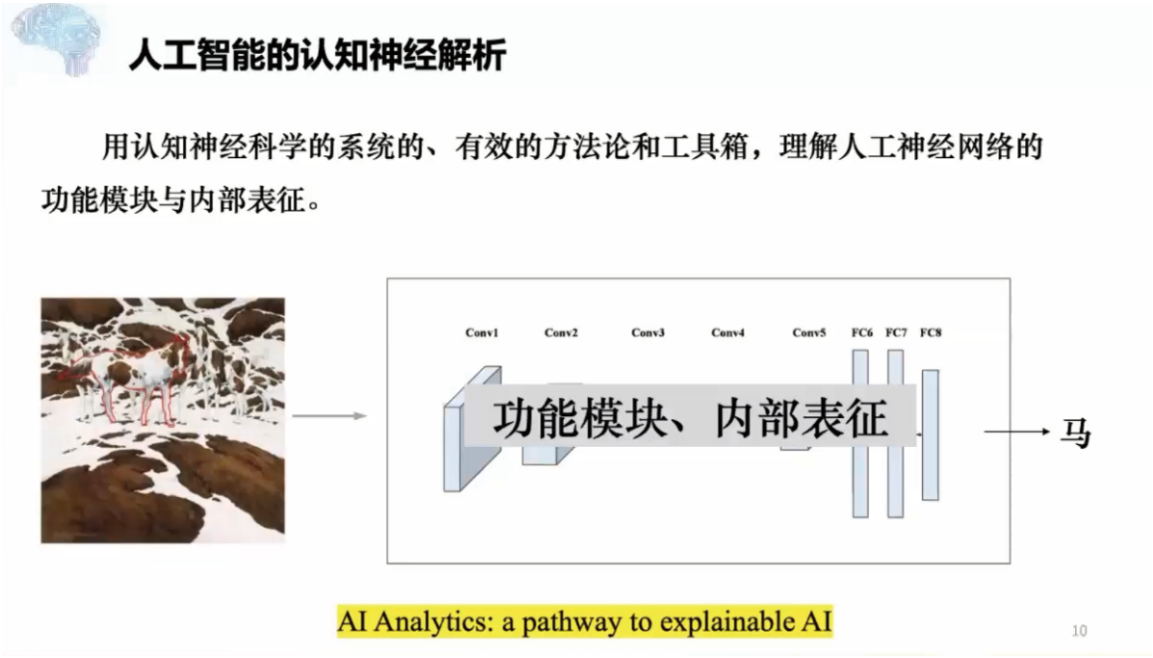


图 5：人工智能的认知神经解析

三、打开深度神经网络的黑箱

3.1 人脑与类脑是否采用了同样的表征来完成任务

图灵测试从本质上来讲，是基于行为主义的逻辑——一个机器只要它在行为上达到人的水平，那么它就具有跟人一样的智能。但是从认知科学的角度，一个更本质的测试应该是：一个智能机器，是否具有与人一样的认知过程。例如，AI 如今能够实现物体识别、目标检测等任务，但是 AI 使用的内部表征和人类是不是一样的？在这个研究，我们将具体回答两个问题：深度神经网络使用什么表征？这种表征和人类相似吗？

我们这里呈现一个性别辨别的任务，下图中左边的是女性，右边的是男性。但是如果我问，你是靠什么特征来进行判断的？他们头发的长短吗？他们的眼睛大小吗？他们脸型的外轮廓吗？还是什么？你可以反省自己到底是靠什么做的判断。

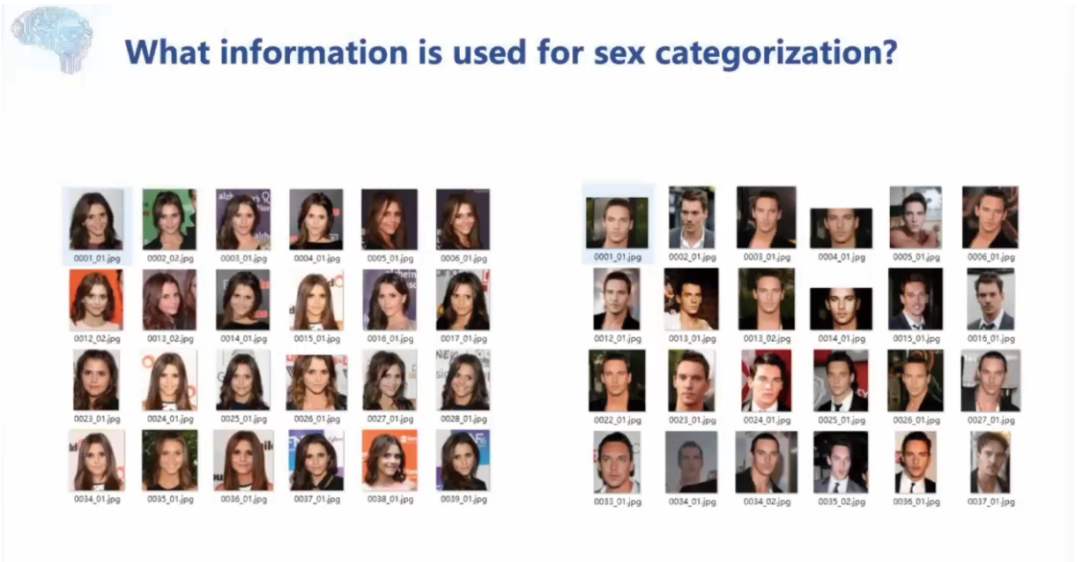


图 6：进行性别分类时所使用的特征信息

你会感受到这个任务很难，**辨别性别很容易，但是理解究竟用哪些特征来做是挺难的**。因为我们进行面孔认知加工是在无意识中完成的，不能被我们意识所觉察到。这里，我们采用认知神经科学的方法，即反向相关的方法（Reverse Correlation），通过结果来回推内部表征。

首先，我们分别把女性面孔和男性面孔取平均，得到女性和男性的平均脸。当我们从女性平均脸平滑的过渡到男性平均脸的时候，大家感受一下效果。

图 7：女性平均脸

这动画给人一个感觉，你对性别的判断类似二分法。开始时是一张女性脸，后面是一张男性脸，中间是感知边界，我们心理的感受并不是随着图像的线形变化而发生线性变化的，而是二分法，前半部分全是女性，后半部分全是男性。这里，我们找到感知边界，生成一张中性脸。

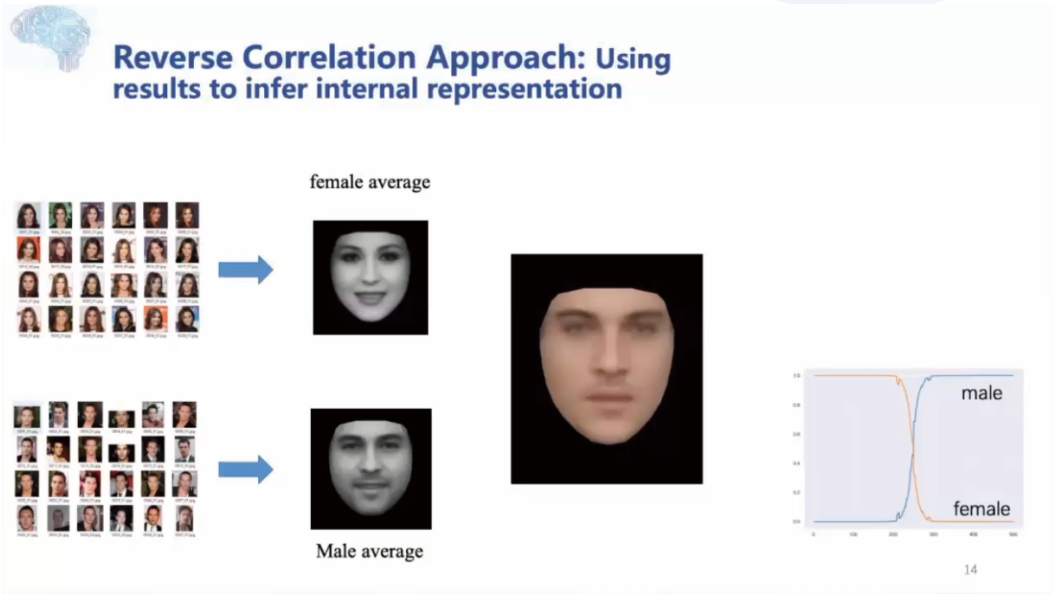


图 8：Reverse Correlation Approach: Using results to infer internal representation

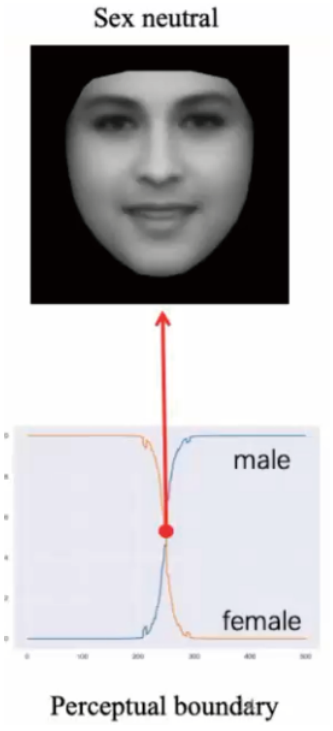


图 9：中性脸示意图

接下来，我们训练一个能识别性别的 VGG-Face 网络。这个网络已经经过预训练，我们只做迁移学习，即把最后一层进行微调，对男性和女性的人脸做识别训练。很快，对性别识别的准确率就达到了百分之百。我们把中性的面孔拿出来加上随机噪音，然后再把这张照片输入 VGG-Face，让它进行分类。添加噪音可以使中性脸被

识别为男性脸或女性脸。我们识别了 2 万张照片，每张照片基底图是一样的，而添加的噪音不一样，这样我们可以得到一组被 VGG–Face 识别为女性的照片和一组识别为男性的照片。

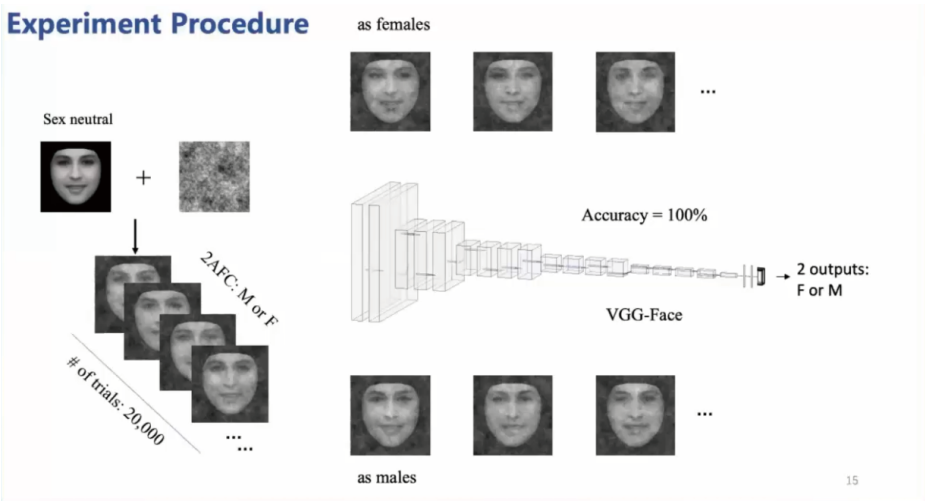


图 10：实验过程示意图

我们把这些照片都贴上了标签，然后把原来的基底图去掉，只留下噪音，并按照性别的标签分别叠加在一起。下图就是 VGG–Face 把面孔识别为女性的面孔特征图。原来的随机噪音看上去无规则，但是通过 reverse correlation 就可以从噪音中提取出结构的信息。我们大致看到，这些信息主要集中在眼睛、鼻子和嘴这些地方，这些特征是 VGG–Face 将面孔判断女性的关键信息。

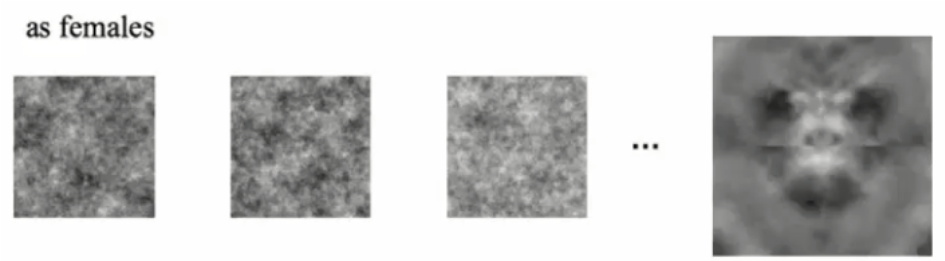


图 11：女性性别判断的关键信息

同样，我们可以把被判断为男性的噪音叠加在一起，得到关于男性的一张特征图。简单对比可以发现，判断为女性的特征图和判断为男性的特征图是不一样的，这两张图的模式很复杂。

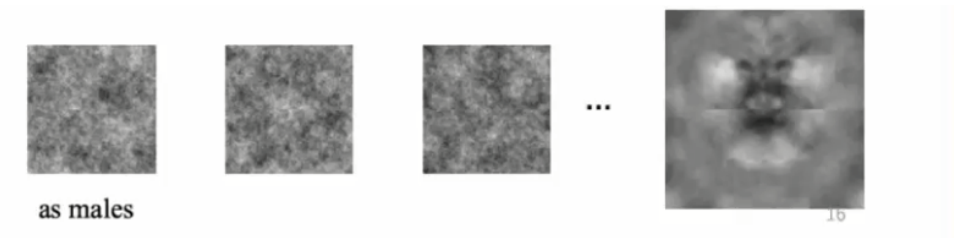


图 12：男性性别判断的关键信息

我们把女性特征图和男性噪音特征图进行相减，得到识别特征图，这张识别特征图就是 VGG–Face 完成性别识别任务的内部表征，它认为这是把男性和女性分开的关键信息。**我们把基底图即中性脸叠加上去，可以看到噪音特征图的极值点大致分布在眼睛和鼻子外侧，以及人中、嘴唇的下沿。**

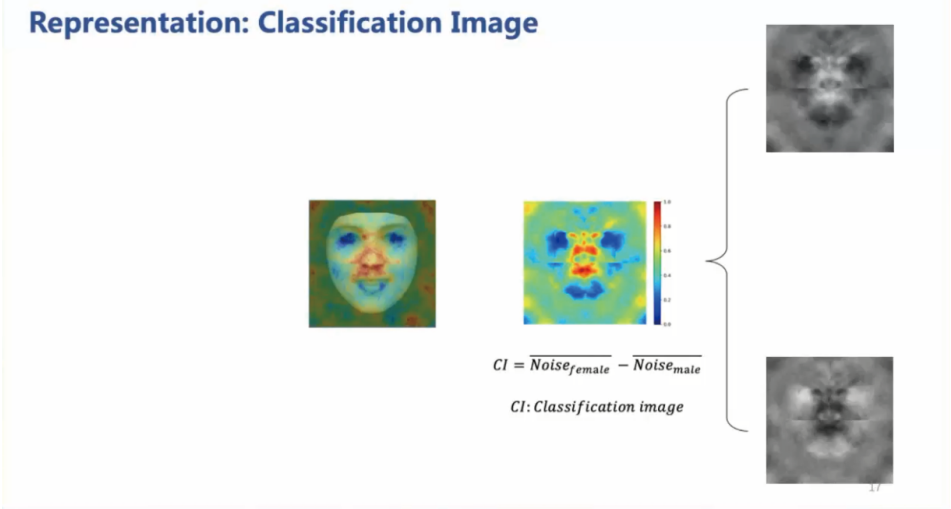


图 13：Representation: Classification Image（1）

我们接下来把这叠加到基底图上，我们就得到了一个标准的男性脸。反之，如果我们把基底图减去这张识别特征图，就会得到一个标准的女性脸。所以我们通过这一系列操作就得到了 VGG–Face 用什么特征来进行性别判断。

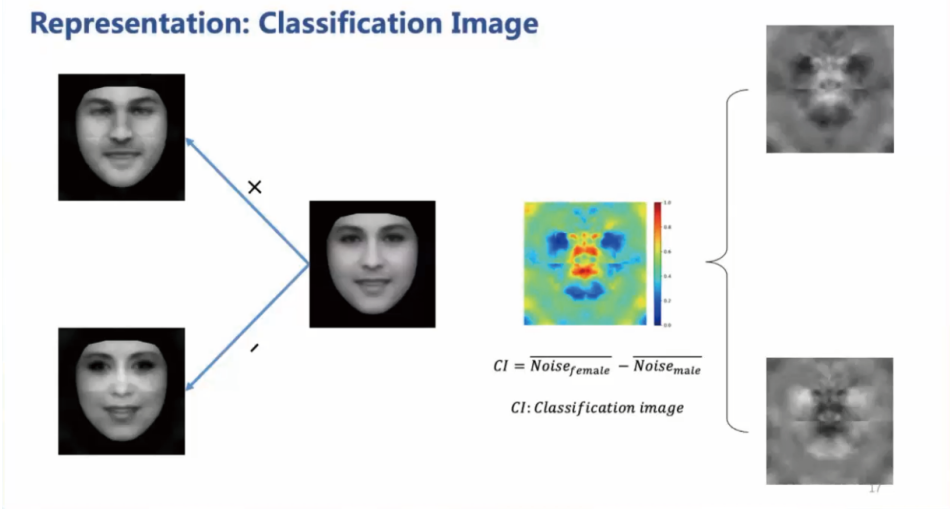


图 14：Representation: Classification Image（2）

如果把 VGG–Face 换成人，结果会如何？我们找人看了这 2 万张图片。在大部分情况之下，被试会说“我怎么知道他是男性还是女性?”我们说“没关系，你猜就是了，跟着感觉走，你觉得它是女性就按 F，觉得是男性就按 M”。于是被试带着困惑、不解和劳累，把这个实验给做完了。这是他们用于区分男性和女性的特征图。我们按照相同的计算，分别得到男性的标准脸和女性的标准脸。

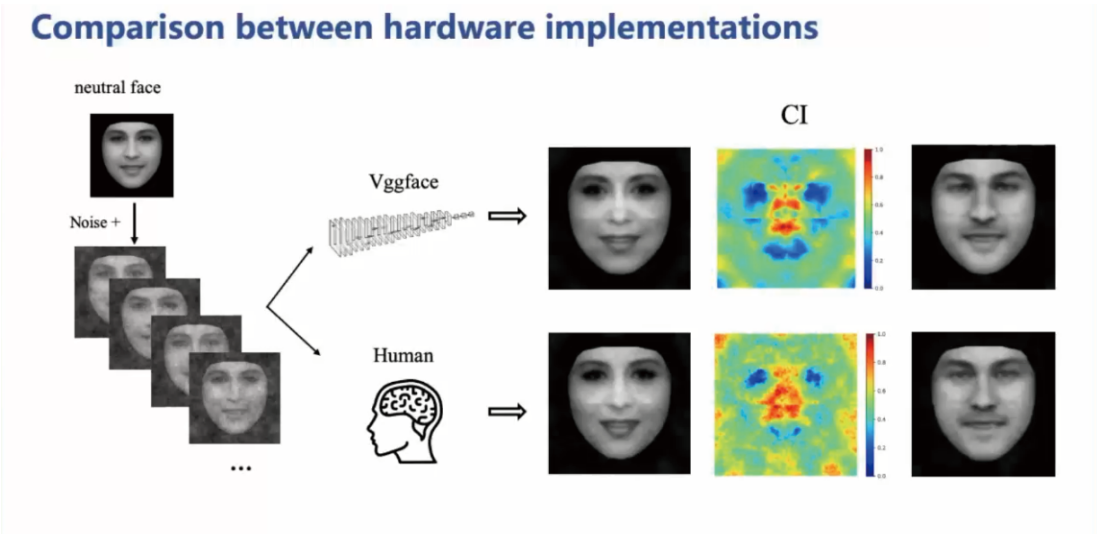


图 15：Comparison between hardware implementations

我们发现在 VGG-Face 的特征图和人类的是非常类似的。事实上，如果我们对这两张特征图计算相关，可以得到 0.73 的相关度。**从这个角度来讲，人类和 VGG-Face 用了类似的表征来完成性别识别的任务**。

进一步，我们来看这个相似是发生在什么空间频率上。在研究中，添加到中性脸的随机噪音是有结构的，由不同空间频率的图组成，下图最左边是低频的，最右边是高频的，我们把低频和高频的信息叠加起来，给大家看到一个实验用的噪音图。

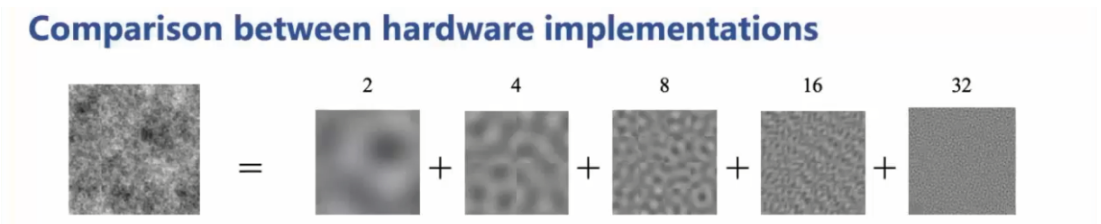


图 16：实验噪声图

现在看在不同的空间频率下面，人和 VGG-FACE 的特征图分别是什么样子。这些特征图也是非常相似的，而且相似度在低频上是最高的，随着空间频率的增加，人和 VGG-Face 的相似度越来越低。所以，**VGG-Face 和人类在完成面孔性别识别任务时，更多依赖于低频的信息**。

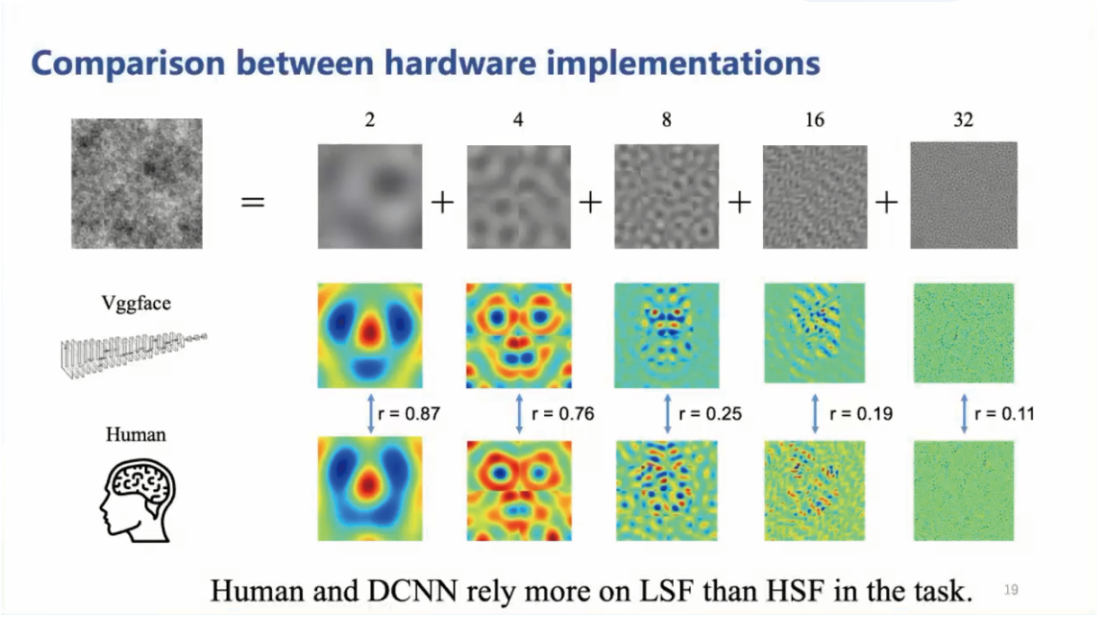


图 17：面孔性别识别时，更多依赖低频的信息

简单总结一下，计算机视觉的奠基人之一 David Marr 提出我们应该从三个层面理解智能。第一个层面是实现的目标或完成的任务，比如这个实验做的就是性别识别任务，这是最高的层面。最低的层面是物理实现的层面，也就是用什么硬件实现。在这个研究里有两种实现的硬件，一个是 VGG-Face，一个是人的大脑，这是两个完全不同的物理层面。

用物理硬件实现目标，中间还需要一个软件的层面，称之为表征和算法。表征和算法在输入和输出之间建立一种转换，这种转换就是智能。智能的本质就是表征。在上述研究里，表征就是把男性和女性区分开的特征图。

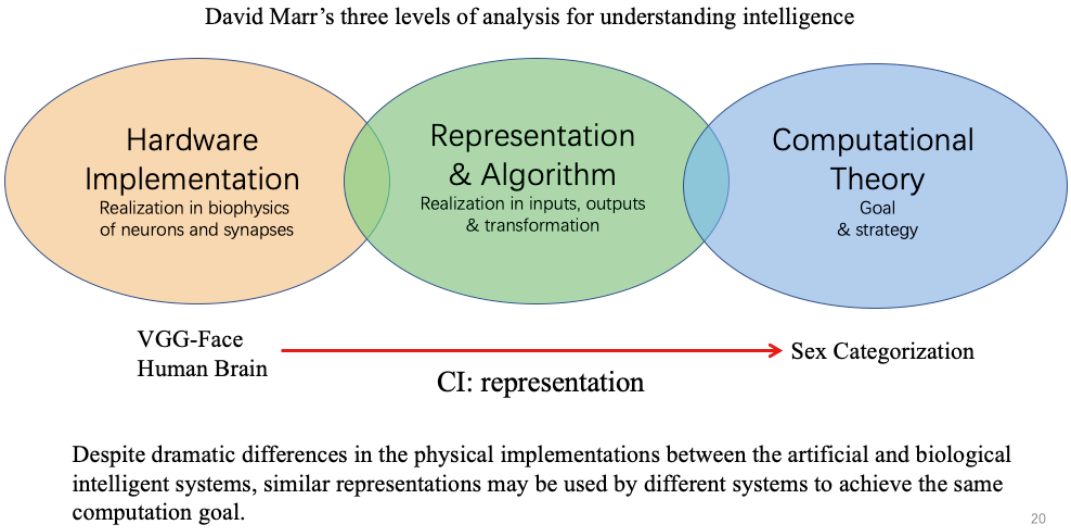


图 18：David Marr 提出我们应该从三个层面理解智能

3.2 类似的任务经验对于形成类似的表征十分重要

VGG–Face 和人类用类似表征来完成性别识别任务，前提条件是什么?

面孔对于人类而言比较特别，我们看到一个面孔，通常需要识别出身份，即直接识别个体，即这是张三。但是对非面孔的物体，我们的识别通常是在类别层面，比如我们看到猫，只会说是一只猫，而不是说这是张三的猫。

其次是对面孔的识别更多依赖低频信息，比如心理学的负片效应，把照片的黑白值翻转，发现识别起来非常困难，同样把低频信息过滤，识别也非常困难。

因为 VGG–Face 是经过面孔识别预训练的任务； 所以，VGG–Face 与人有类似的表征，可能是因为上述这两个原因，即：（1）VGG–Face 和人都是在个体层面上识别物体；（2）VGG–Face 和人因为处理过大量的面孔，因此会对面孔的独特特征（如低频信息）敏感。

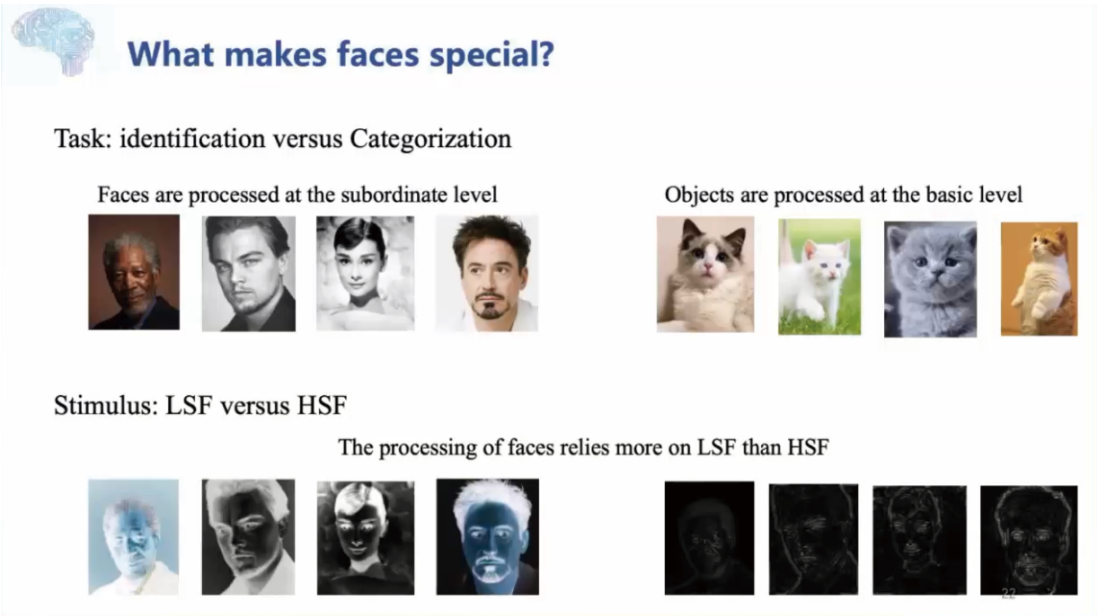


图 19：什么让面孔变得独特?

首先，我们来验证第一个可能性： 共同的任务经验。这里，我们选择 AlexNet。AlexNet 也是预训练网络，它不做面孔识别而做物体分类，我们把最后一层微调，让它做识别男性和女性的分类任务，正确率 93%。即，虽然 AlexNet 是用来训练物体分类的，但是也能够把男性和女性区分，正确率也相当高。

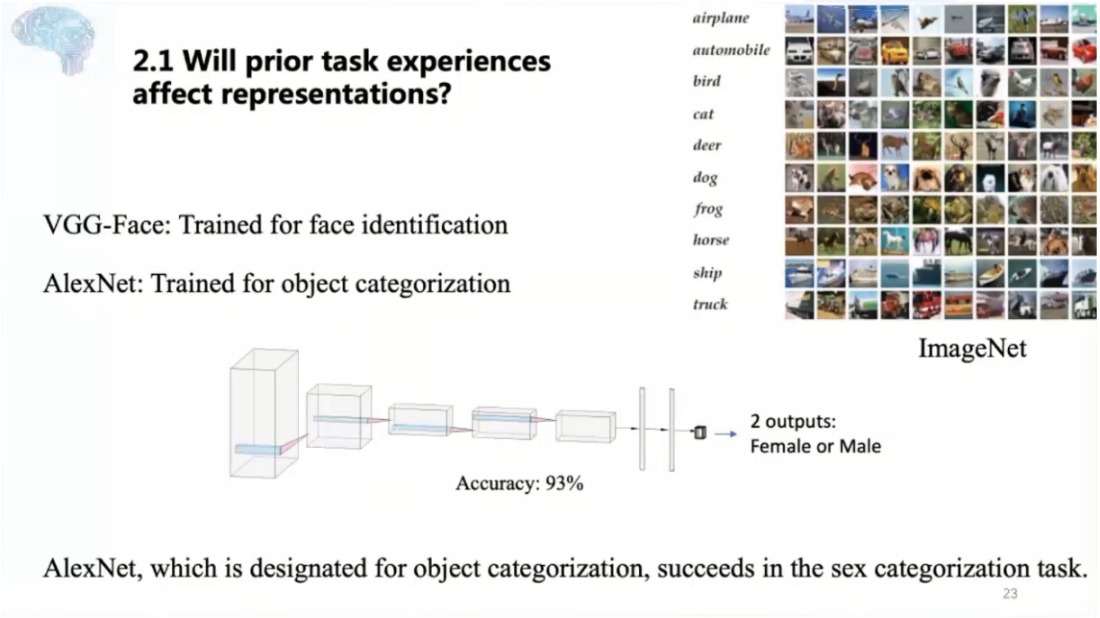


图 20：Will prior task experiences affect representations

现在问一个有趣的问题，AlexNet 在性别辨认上也能达到和人一样的准确度，但是 AlexNet 用的是和人类似的表征吗？我们来看 AlexNet 辨别男性和女性的特征图，如下图所示，肉眼能够辨别两者存在非常大差别，基本不相关，相关度等于 –0.04。我们把它叠加到原来的基底图上去，得到的人脸也没有明显的性别特征。所以从这个角度来讲，我们发现 **AlexNet 虽然能够区分男性和女性，但是它所用的表征是完全不一样的**。

我们做进一步的空间频率分析，把噪音特征图分为不同的空间频率，可以看到，基本上 **AlexNet 和人类的各频率的噪音特征图是不相关的**。

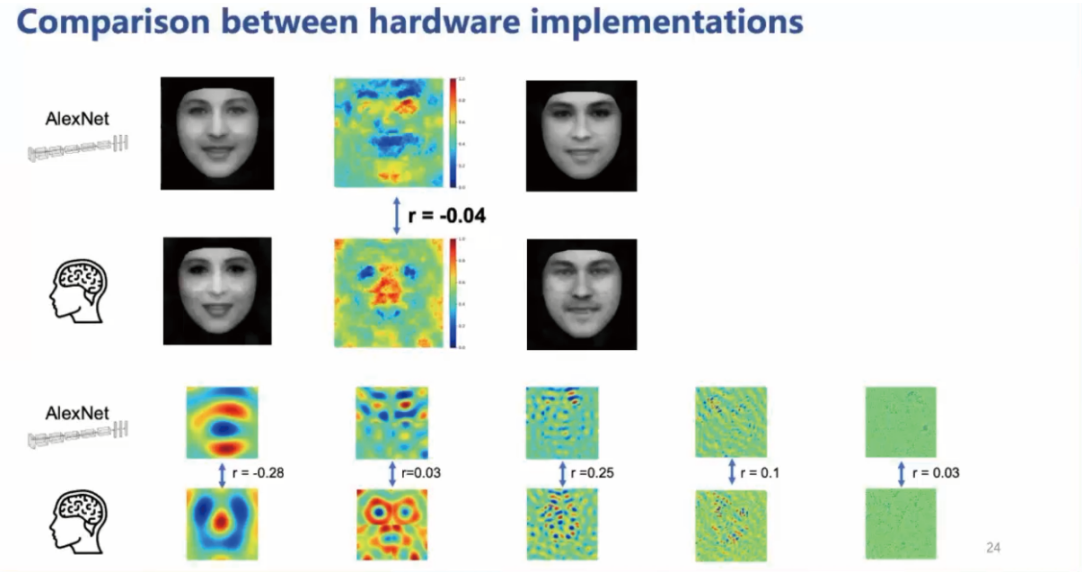


图 21：AlexNet 区分性别所用的表征与人类是完全不一样的

回到实验的第一部分结论，我们发现预训练任务非常重要。为什么 VGG-Face 和人类在区分男性女性时用的表征是相似的？因为它们都被训练在个体层面上进行加工，而 AlexNet 是在类的层面上进行加工，从这个角度来讲，导致它们使用呢不同的表征。

这一点我们可以从进化的角度来理解。我们之所以从单细胞变成现在多细胞的动物，就是因为我们在不断地完成大自然交给我们的任务；一旦完成不了，那只有一个结果，就是基因被淘汰。也就是说，we are what we do。我们的智能是我们过去所完成的任务所决定的。

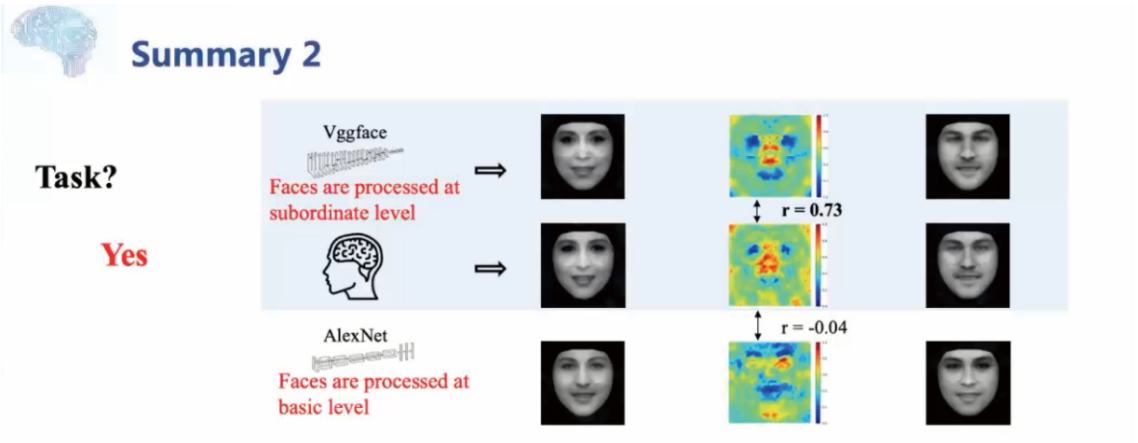


图 22：我们的智能是我们过去所完成的任务所决定的

总结一下：认知神经科学发展了一系列有用的工具和方法论以及实验范式，这些范式有助于我们了解深度神经网络内部特征和模块，得到可解释、可预测的深度神经网络。更进一步，认知科学、神经科学和智能科学的深度交叉所形成的认知神经智能科学将会为揭示智能的本质，提供一个新的视角。具体而言，一个理想的研究智能模式是：通过神经科学发现一个大脑工作的机理（brain inspiration），根据认知科学来来对该机理进行建模（cognitive modeling），然后用计算科学来开发一个计算复杂度适度的算法（physical implementation）来解决一个真实的现实问题。

北大教授吴思：生物视觉和计算机视觉之间的对话

转载自：AI 科技评论

6 月 22 日，北京智源大会举行了认知神经基础专题论坛，来自北京师范大学认知神经科学与学习国家重点实验室的毕彦超教授、北京大学心理与认知学院的方方教授、清华大学心理学系的刘嘉教授、北京大学计算机系的吴思教授、中国科学院自动化研究所的余山教授分别做了报告，共同探究认知神经科学能为 AI 带来什么启发。

第四位报告者是北京大学计算机系的吴思教授，演讲题目为《生物视觉和计算机视觉之间的对话》。在报告中，吴思教授指出，生物的视觉识别机制和深度神经网络的图像识别机制有非常大的区别，生物的视觉识别涉及自上而下通路和自下而上通路的交互，而深度神经网络只模拟了第二种通路。自上而下的视觉通路涉及生物视觉感知的全局性、拓扑性、多解性等特点，尤其是理解图像时会面临数学上的无穷解问题，而这些特点或许就是深度神经网络下一步的改进方向。

以下是演讲全文。

我的报告内容是生物视觉和计算机视觉研究的彼此影响，以此说明神经科学和人工智能研究的互动关系。这两个领域本质上都是在解开智能的黑箱，所以两者之间相互启发是非常自然的事情。

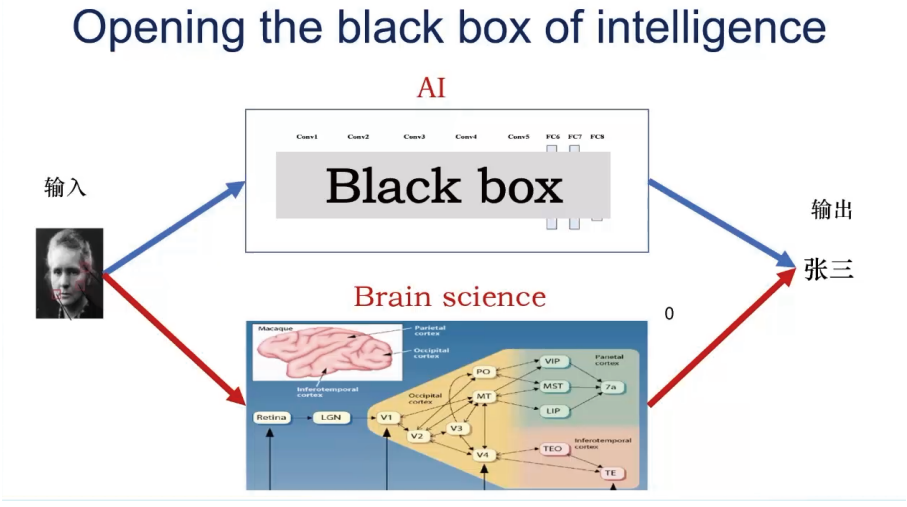


图 1：打开人工智能的黑箱

一、深度神经网络只模拟了部分生物视觉

深度神经网络是近年来人工智能兴起的引擎，已经非常成功，在一些大型数据集对物体的识别率甚至超过人类。但是，深度神经网络还面临很多问题。

第一，**深度神经网络更多是模拟了大脑视皮层中的前馈、层级结构信息处理的方式**。但是大脑的视觉系统比这复杂得多，所以在很多行为上人脑和深度神经网络有非常大的不同。在很多任务上，人的表现更加高明。

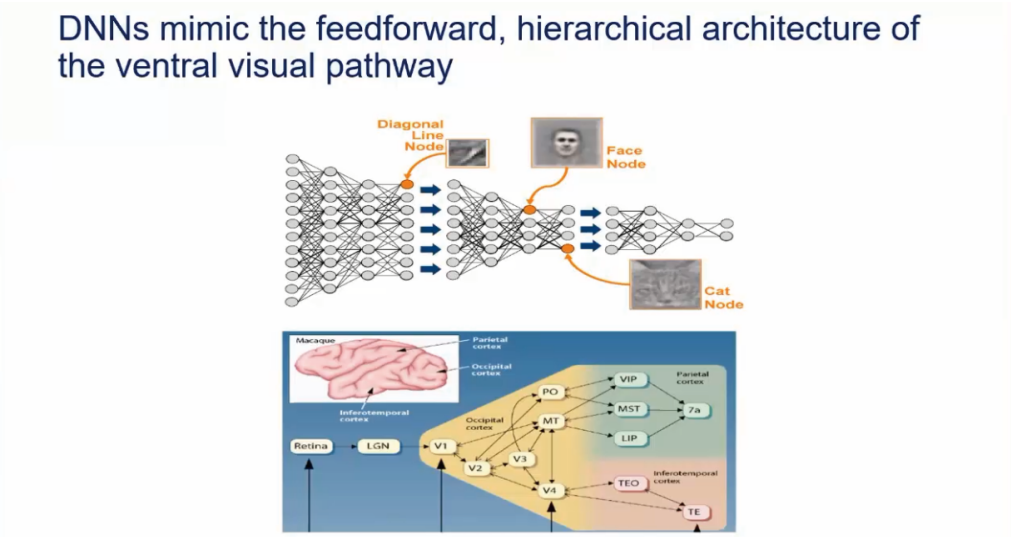


图 2：深度神经网络只模拟了部分生物视觉

举个简单的例子。如下图所示，左边是一头熊，熊的局部信息被去除了，只剩下轮廓，而我们人类一眼就能认出这是一头熊。而右边的图则是把熊分成小块然后打乱，只保留局部的信息，全局信息则没有了。我们可以发现这些小块包含熊的眼睛、嘴巴、身体，但是很难认可右边的图是一头熊，深度神经网络却一眼认出右边的图是一头熊。

通过对比可以发现，深度学习网络的物体识别机制和人类有很大不同。**人类能够获取物体的全局信息进行识别，而目前深度神经网络只能利用局部信息进行识别。**

图 3：深度神经网络只是利用局部信息进行识别

无法获取全局信息是深度学习特别是前馈神经网络面临的一个基本问题，这个基本问题其实很早就被意识到了。人工智能的先驱 Marvin Minsky 在 1969 年就指出，前馈神经网络很难做拓扑性质的识别。

拓扑学是研究几何图形或空间在连续改变形状后还能保持不变的一些性质的学科。它只考虑物体间的位置关系而不考虑它们的形状和大小。在拓扑学里，重要的拓扑性质包括连通性与紧致性。

全局信息很难用前馈网络获取，即使要获取其计算复杂度也呈指数增长。拓扑信息和全局信息的获取是深度学习网络面临的基本问题。

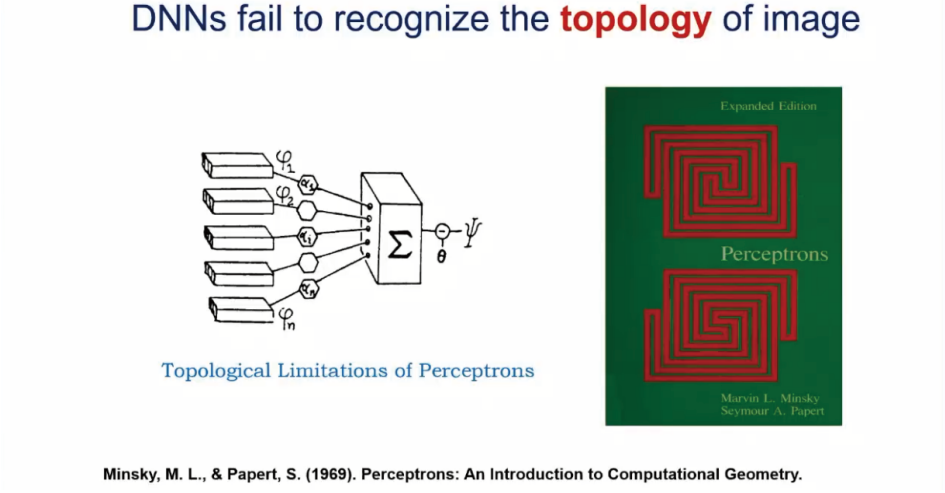


图 4：DNNs fail to recognize the topology of image

所以，我们有必要理解生物视觉系统如何获取全局信息。神经科学领域一直有一个广泛争论，就是人类识别物体到底是根据全局信息还是局部信息。这两种观点对应的典型例子是两种画派，如下图所示，左边的画属于印象主义，如果只看局部的话是看不清眼睛或鼻子的，但是只要从整体进行识别就能知道这是个男人，这是从全局信息进行物体识别的例子。右边的画属于立体主义，这幅画把每个局部信息特别放大，毕加索说画中是一位美丽少女，但是很多人都认为看不出来，因为不能用局部信息拼成整体信息，这是从局部信息进行物体识别的例子。

图 5：物体识别中的整体和局部信息

深度学习网络是通过聚合局部信息逐步构建复杂信息来识别物体的，相反，在认知神经科学领域有一个理论叫“逆向层次论”，这个理论指出，**人类对物体的识别是从简单到复杂、从整体到局部**。

“逆向层次论”和我们的生活经验相一致，如果一个人在我们视野中一晃而过，你马上会反应到这是个人，然后再识别对方的身份，这就是一种从整体到细节的识别过程。

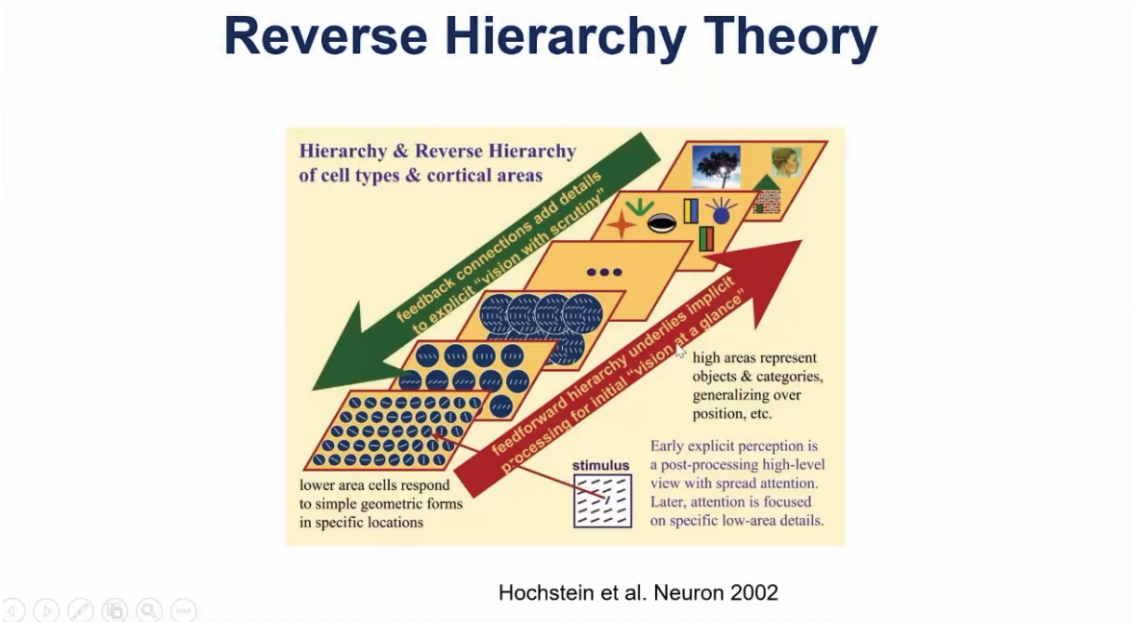


图 6：逆向层次论

我们从神经科学的角度来看人类视觉认知与机器学习的一个重大不同点。下图展示了一个实验，被试是盲视。盲视是指，意识层面“看不见”物体但却能“感知”到物体的存在。

图 7：盲视实验

大量实验表明，人类要看到或意识到物体，需要物体信息至少在视觉皮层 V1 中被接受到。假设 V1 受到损伤，就可能会产生盲视现象。这时还能感知到物体是因为皮层下通路还存在，皮层下通路是从视网膜直达上丘然后再到高级皮层的一条短路径。

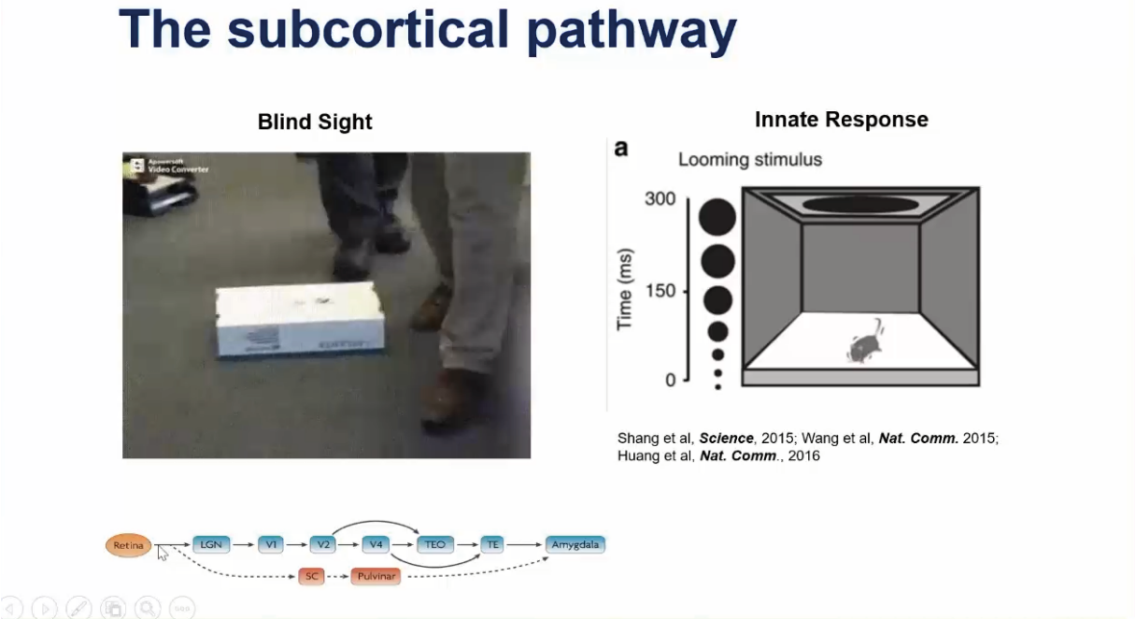


图 8：The subcortical pathway

科学家利用动物实验更好的证明了这一点。他们把老鼠放在笼子里，天花板上会呈现一个动态刺激，即一个小的光斑很快变大，这模仿了在自然环境中老鹰向老鼠俯冲下来时，老鼠视网膜接受到的光信号。这时候，老鼠本能的第一反应是装死。科学家发现，在上丘处通过操纵神经元反应可以让老鼠看到运动光斑后不再装死，或者即使没有运动光斑的出现老鼠都主动装死。这个实验表明**本能的快速反应走皮层下通路，而没有走深度神经网络模拟的皮层上通路**。

在上述老鼠将运动光斑当成老鹰的实验中，老鼠根本没有刻意去识别刺激是光斑还是老鹰，立刻装死。这是动物的本能反应，即老鼠没有做细节的特征提取也能识别运动模式。

我们参考这个例子，提出了一种新算法，在识别运动模式时不做特征提取。我们建立了一个模型，这个模型包含两个部分，下图左下方是外界输入，黑色圆圈中的网络表示“视网膜”。这里“视网膜”的计算很简单，它把运动模式投射到高维空间，使运动模式变成线性可分的，然后再输入到抉择网络。“视网膜”的神经元特别多，相当于一个库网络。我们不需要训练库网络和抉择网络，只需要训练库网络和抉择网络之间的连接。

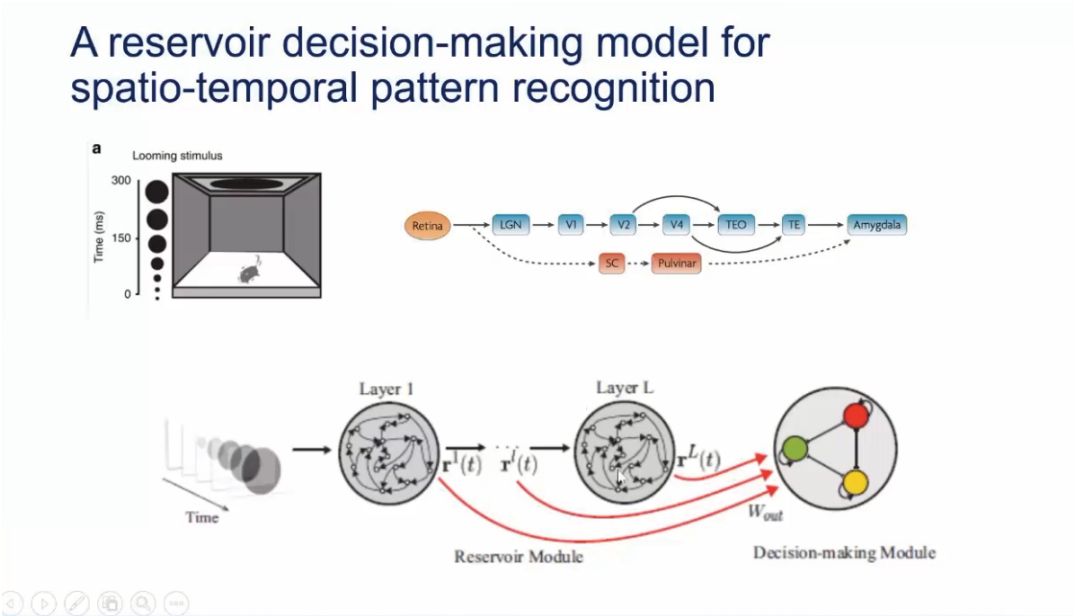


图 9：识别运动模式的算法模型

关于抉择网络，我用两个神经元来举例解释一下，如下图所示，每个抉择神经元代表要识别的一类运动模式。这些神经元的动力学特别的慢，因为要识别运动模式，关键是要抓住输入的时间结构，不仅仅是空间结构。这些抉择神经元之间存在相互抑制，每个神经元通过库网络输入收集证据，如果证据支持自己编码的运动模式，这个神经元的反应就会抑制其它神经元的活动而最终胜出。

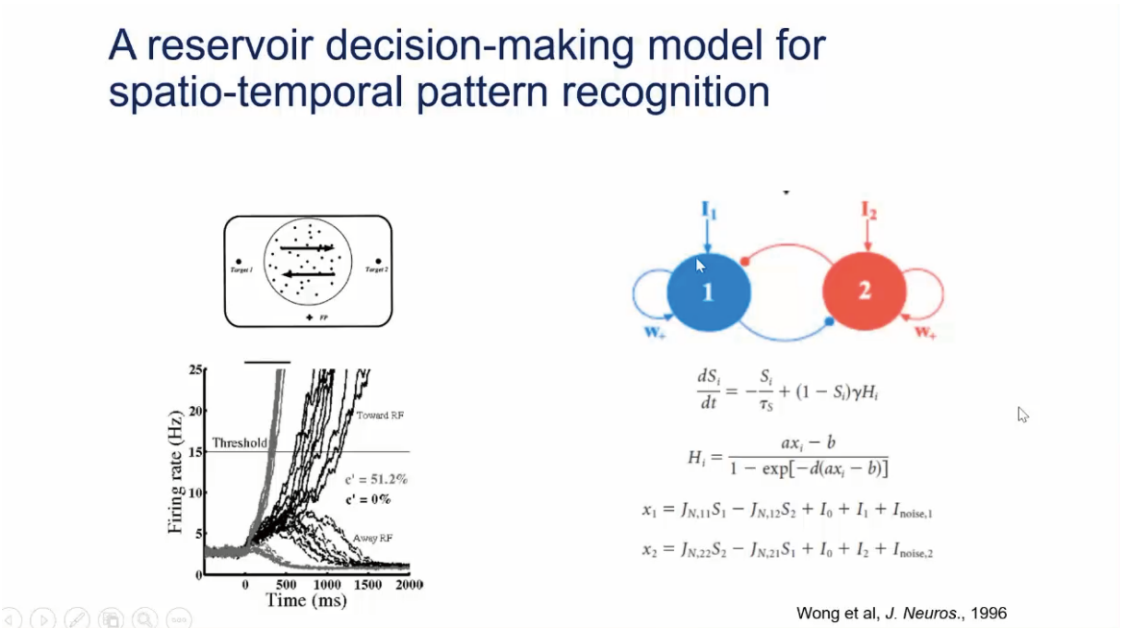
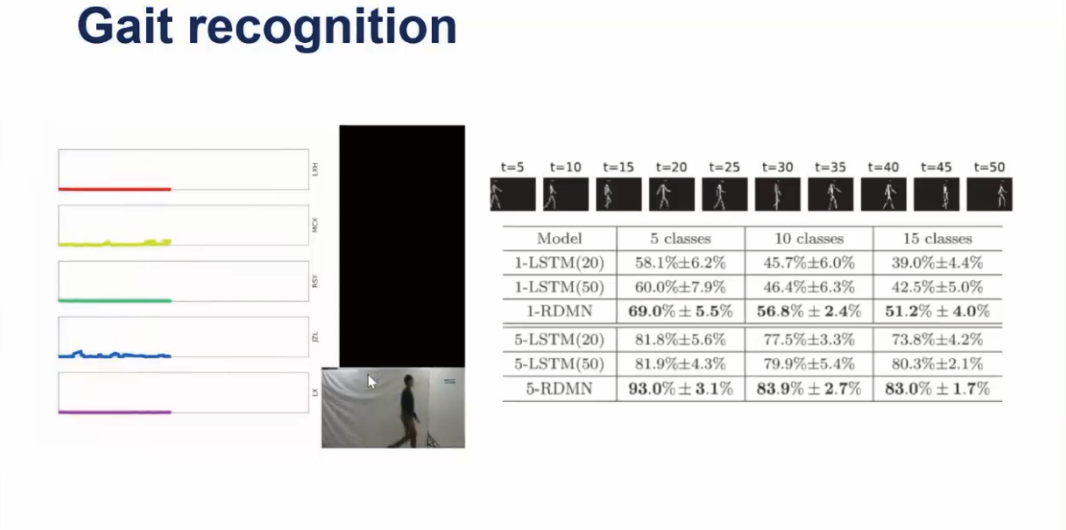


图 10：A reservoir decision–making model for spatio–temporal pattern recognition

这个模型的计算本质是时空模式的识别，所以我们可以把这个模型推广，用来做步态识别。在这个任务中，人在屏幕前走 1–2 回，然后把步态输入到模型中，进行识别。这个模型的优点是可以小样本训练，只需要 1–2 回的数据就能马上学会一个人的步态特点。



Model	5 classes	10 classes	15 classes
1-LSTM(20)	58.1%±6.2%	45.7%±6.0%	39.0%±4.4%
1-LSTM(50)	60.0%±7.9%	46.4%±6.3%	42.5%±5.0%
1-RDMN	**69.0% ± 5.5%**	**56.8% ± 2.4%**	**51.2% ± 4.0%**
5-LSTM(20)	81.8%±5.6%	77.5%±3.3%	73.8%±4.2%
5-LSTM(50)	81.9%±4.3%	79.9%±5.4%	80.3%±2.1%
5-RDMN	**93.0% ± 3.1%**	**83.9% ± 2.7%**	**83.0% ± 1.7%**

图 11：Gait recognition

二、生物视觉是一个动态交互的过程

我们介绍一个心理物理实验来展示由整体到局部的识别实际上是不可避免的。请大家看下图中呈现的图像，猜一猜是什么。

图 12：图中呈现的图像是什么?

如果你过去没有见过这张图的话是肯定猜不出来的，所以我把图像的轮廓画出来。

图 13：画出轮廓后可以识别出这是一头牛

现在你就能看出来图中是一头牛。如果把牛的轮廓去掉，你还是觉得图中是一头牛，因为这时你大脑中已经有了自上而下的牛的先验知识。但这只是其中一个答案。我也可以画一只手的轮廓，然后轮廓去掉，这时候你又会觉得图中是一只手，因为你有了自上而下的手的先验知识。

图 14：画出手的轮廓

我还可以在图中画一条鱼，我相信这时候你又会觉得图中是一条鱼。

图 15：画出鱼的轮廓

这个实验表明人类识别物体时，大脑皮层的自上而下的信号非常重要。

这个简单实验揭示了图像理解的一个深刻数学问题，**即给定一副图像，它的解释理论上有无穷多个。**注意图像理解跟物体识别不一样，图像理解涉及两个基本操作，一个是图像分割，一个是物体识别。

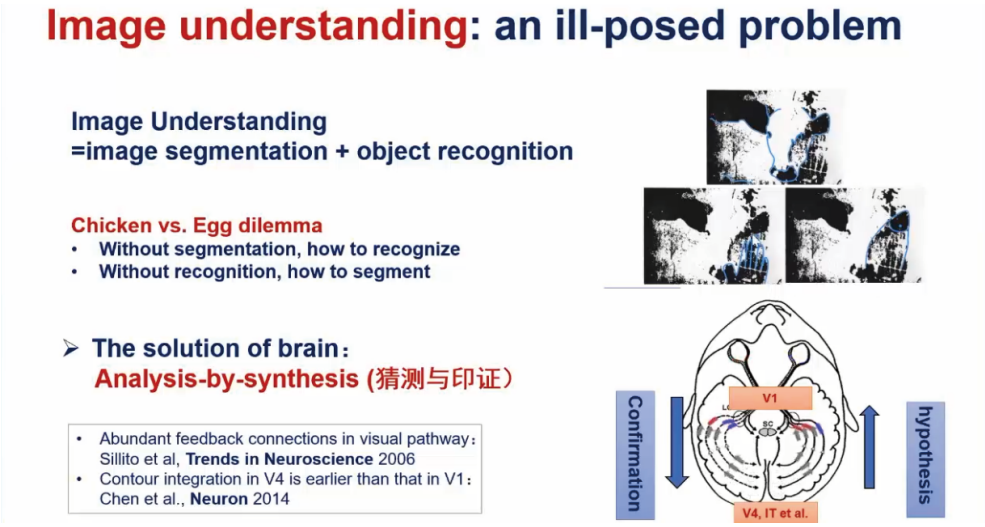


图 16：Image understanding: an ill-posed problem

但两者的顺序是一个鸡生蛋或蛋生鸡的难悖论：给你一幅图像，没有合适的分割，如何做好识别；但另一方面，如果没有预先识别物体，又如何做合适的分割呢？从数学上来说，一幅图像有无穷多的分割和识别的方式，所以在数学上这是一个不适定的问题。无论是人类还是 AI，图像理解时都面临这样的难题。

大脑解决这个问题的思路是一个“猜测与印证”的过程。当我们识别物体时，物体的图像信息快速传递到高级皮层，即通过所谓的快速通路，在高级皮层做出猜测。猜测结果再通过反馈连接，和新的输入交叉印证，如此反复进行后，才能识别物体。

我们在日常生活中很难意识到这个过程，因为在日常生活中，很多时候只需要一两个回合就能成功识别。但的确有的时候一个图像看得不太清楚，我们会盯着它左看右看，大脑内部可能就进行了信息的上传、下传的交替，不断地进行“猜测 – 印证 – 猜测 – 印证”，只要印证结果是否定的，这个过程就会一直进行下去，直到得到肯定的结果。

神经生物学充分证明人类大脑的识别机制确实如此。从解剖上来说，从高级视皮层到初级视皮层的反馈连接比前馈连接还要多，相比之下深度学习网络主要考虑的是前馈连接。电生理实验证据也表明，大脑对物体的识别先发生在高级视皮层，然后才发生在低级视皮层。

总的说来，生物视觉识别至少有两条通路，快速的通路对物体整体进行识别，其结果帮助慢速通路对物体局部信息的识别。

下面以我们最近的一个工作来介绍整体识别可能如何通过反馈提高局部识别。我们考虑对物体进行识别时，先对物体大类识别，然后根据大类信息帮助进行小类识别。比如我们看到一个图片，先识别这是动物，再识别这是猫，还可以进一步识别这是什么品种的猫。我们发现大类信息可以通过先正后负的反馈信息帮助小类信息识别。

第一步是正反馈（Push feedback），其作用是压制类间的噪音。假设高级脑区识别出物体是一只猫，就告诉低级脑区不要再处理狗的信息了。这是正反馈，增强猫的信息，压制狗的信息。第二步是负反馈（Pull feedback），其作用是压制类内的噪音，即在猫的信息中把猫共性平均值减去，把不同猫之间的细微差别放大。

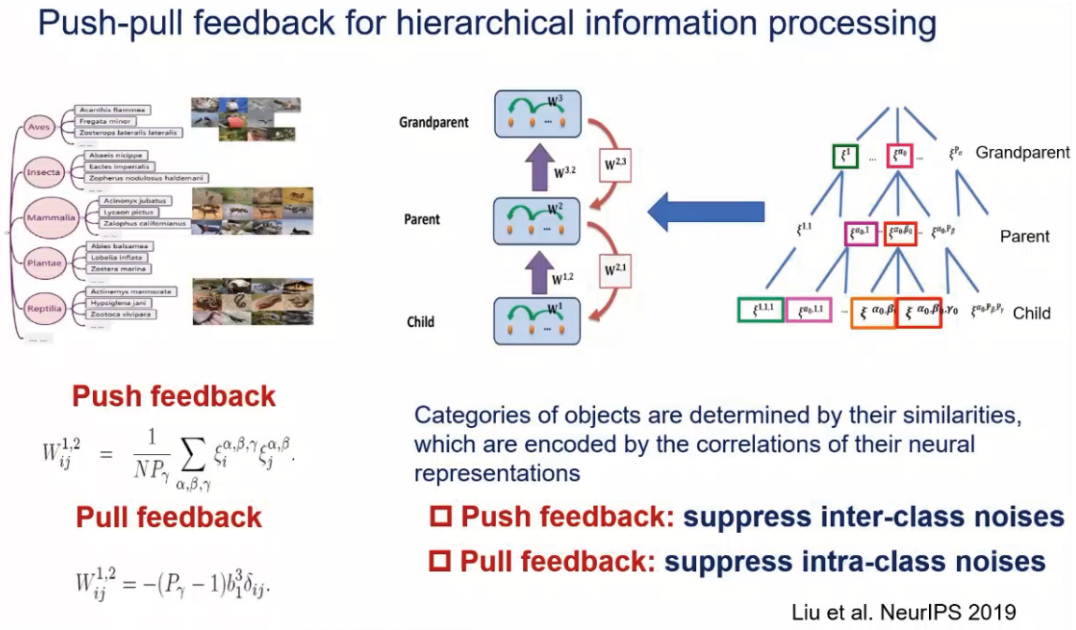


图 17：Push–pull feedback for hierarchical information processing

总的说来，生物视觉的识别机制和深度神经网络的图像识别机制有非常大的区别，生物的视觉识别涉及自上而下通路和自下而上通路的交互，而深度神经网络只模拟了第二种通路。自上而下的视觉通路涉及生物视觉感知的全局性、拓扑性和多解性等特点，而这或许就是深度神经网络下一步的改进方向。认知神经科学和人工智能应该多互相对话、互相借鉴，按照过去的经验，这样做经常能带来惊喜。

中科院研究员余山：从脑网络到类脑计算

整理：AI 科技评论

他山之石，可以攻玉。对于人工智能研究，脑科学无异是最重要的「他山之石」了。

近年来，人工智能在经历过一波由深度学习带来的火爆之后，已然进入深水区；如何通向强人工智能，逐渐成为智能研究的各界人士共同关注的中心话题。「类脑计算」正是智能研究人员尝试以脑科学之「石」攻智能之「玉」的重要方向。

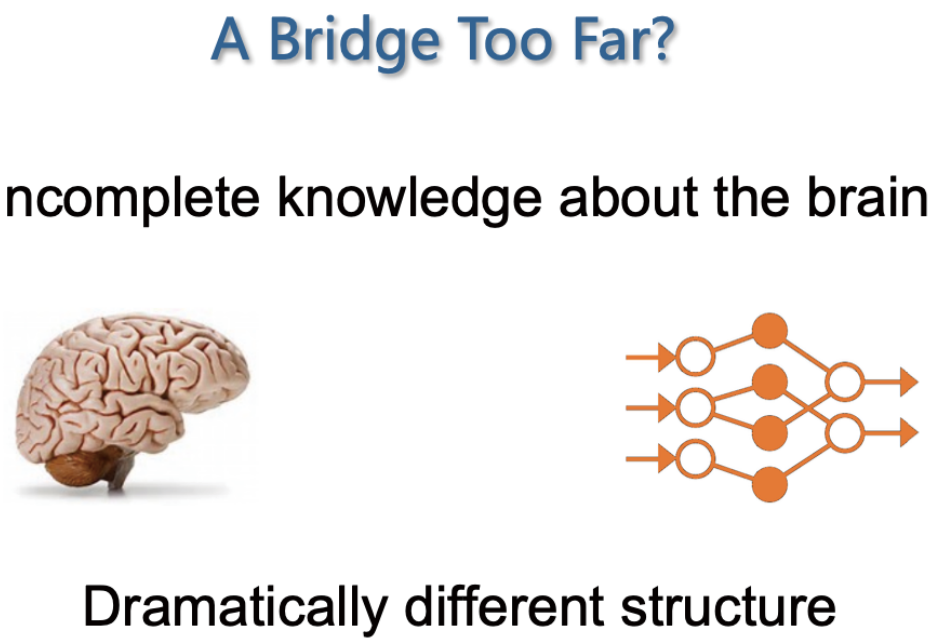


图 1：对于大脑我们仍有很多未知

6 月 22 日，在第二届智源大会“认知神经基础专题论坛”上，中国科学院自动化研究所余山研究员作了“From Brain Network to Brain-like Computation”主题报告。余山研究员借鉴 Marr 对视觉体系的划分，将类脑计算的研究分为四个层面：硬件、算法、计算、学习。针对每一层面，余山研究员做了或简或详的介绍，颇具启发性。

图 2：Marr 将类脑计算的研究分为四个层面

余山研究员认为，尽管当前人类对大脑的认知并不充分，但这并不阻碍智能研究的各界人士去借鉴已有的神经科学和脑科学的知识，从而来发展对智能系统的研究和设计。

一、硬件层面：存算一体设计结构

传统计算机使用的是冯诺依曼架构，其基本架构包括控制器、运算器、记忆单元、输入系统和输出系统等五个组成部分；其中控制器和运算器构成了处理单元（CPU）。

做数据处理时，计算机把数据从存储单元调到处理单元，运算之后再返回到存储单元。但这种操作方式，会导致存储单元和处理单元之间进行非常高频的数据搬运，从而带来极高的能耗。

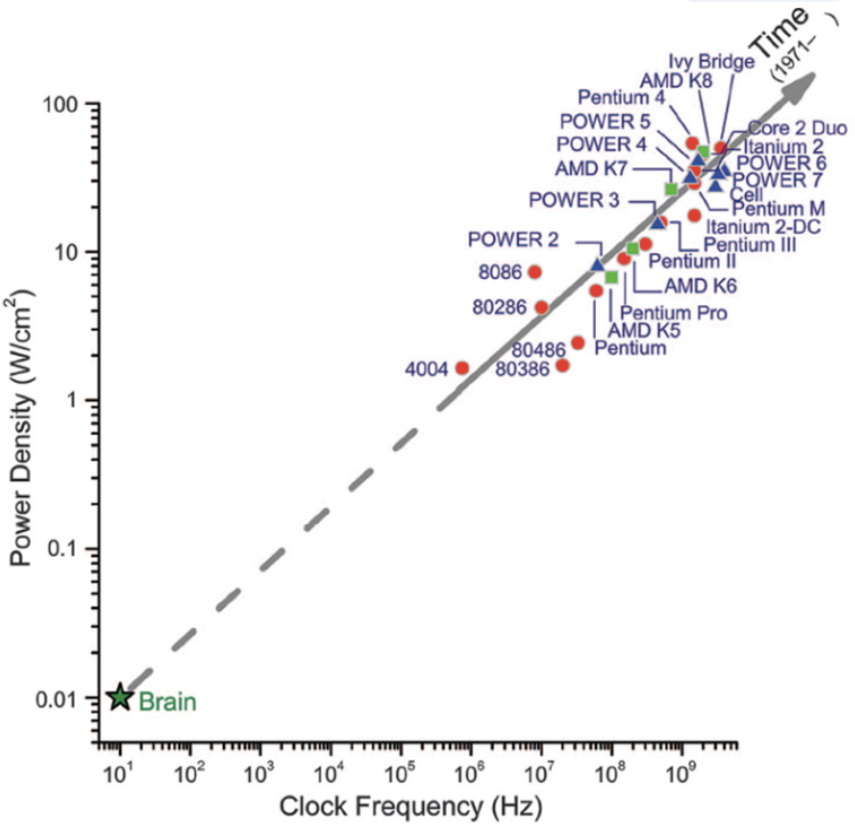


图 3: 计算 频率与能量密度的关系 (时钟频率代表数据在存储单元和处理单元之间调用的速度，能量频率代表功率)

近年来计算机迎来了高速发展，GPU 时钟频率不断提升，但也带来了能量密度逐年提升的问题。以 IBM 在 2000 年开发的一个用来做生物信息学研究的计算机为例，其包含了 144TB 的内存，14 万个处理器，功耗高达 1.4 兆瓦。每当这台计算机运行时，就必须有一个专门的电站为其供能。

反过来，我们看人脑，具有如此高的智能，然而其功耗却只有 20 瓦左右，仅相当于一颗黯淡的白炽灯的能耗。如此大的差别，原因是什么呢？原因自然很多，但重要的一点是，不同于冯诺依曼机，人脑的计算是“存算一体”。在人脑的神经网络中，信息的存储和处理并不分开，神经网络本身即是存储器，又是处理器。

借鉴人脑的这种特点，近年来，有越来越多的研究团队加入了“存算一体”芯片研制中，其中 IBM 研制的 TrueNorth 和清华大学研制的 Tianjic 是这方面最出色的代表。这种芯片被称为神经形态或神经拟态芯片，极大地解决了数据频繁搬运所带来的能耗问题。

二、算法层面：借助突触式信号传递

在算法层面，生物脑和人工神经网络之间具有非常重要联系。余山研究员在报告中提了两个例子。第一个例子是突触的概率释放与 Dropout 算法之间的关系。

在生物神经网络中，神经元之间的连接是通过一个叫做突触的结构进行的，这个结构也是两个神经元之间进行信息交互的地方。当前神经元有一个动作电位时，它会释放某种神经递质，这种递质被后神经元吸收之后便会

转化为电信号，从而实现电信号在神经元之间的传递。

在两个神经元之间信息传递的关键是： 电信号促使化学物质释放。这种方式存在缺点，即神经冲动导致神经递质释放并不总是成功——成功概率的中位数仅在 0.2~0.3 之间，即有 80% 左右的概率会出现信息传输失败。然而，如此低的成功率却有它独特的意义。低成功率，可以使神经网络更快、更好地学习。

人工神经网络训练方法 Dropout 正是对这种现象最好的借鉴： 在网络训练时，随机关闭某些神经元； 而在测试时，让所有神经元都工作。结果显示，利用这种方法，神经网络的学习能力将有明显地提高。第二个例子是有关神经网络的临界状态。我们先介绍一个概念： 神经元的传播系数。简单来理解，即一个神经元能够激活的神经元个数。

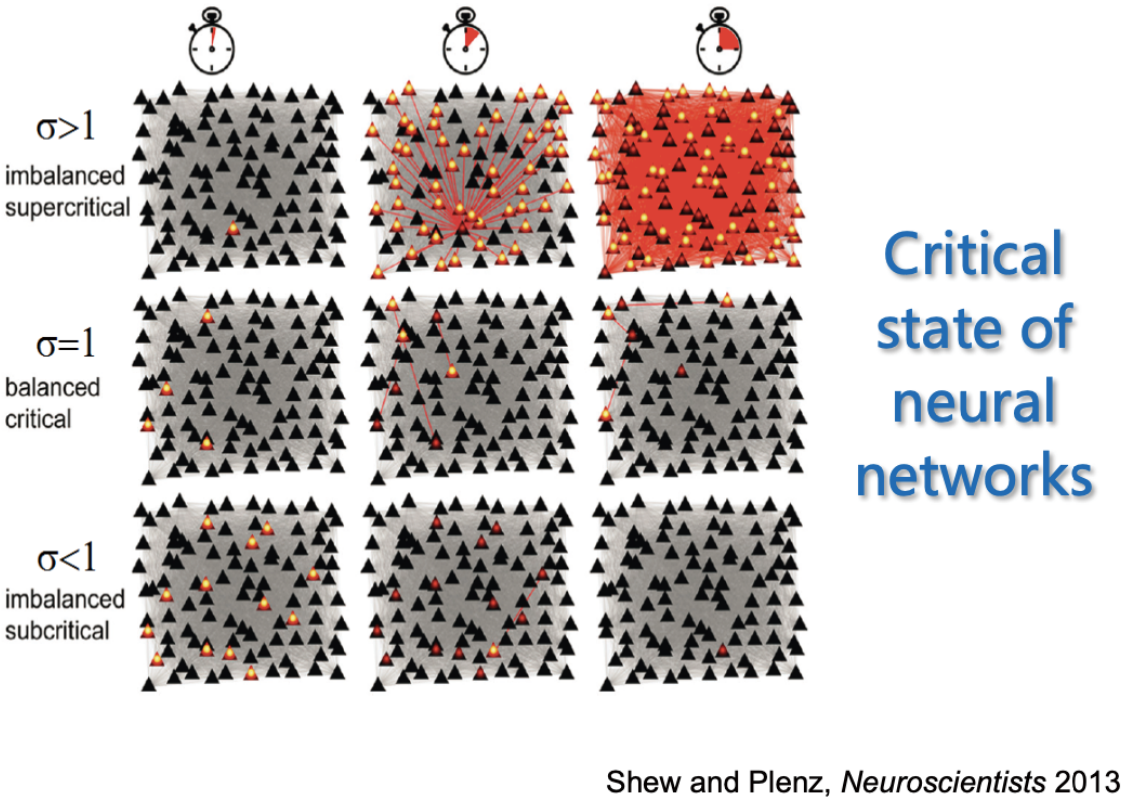
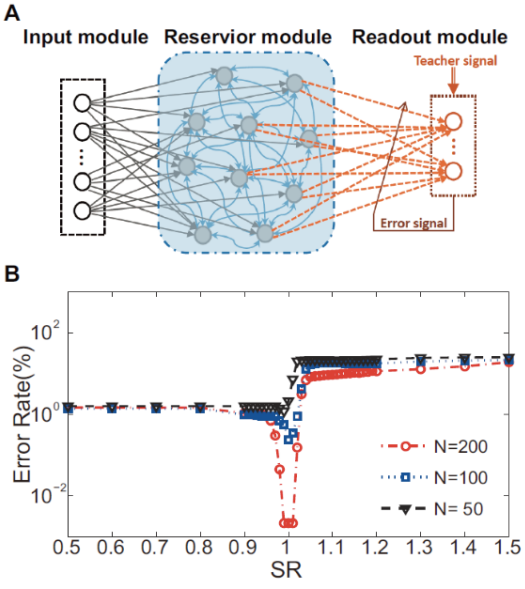


图 4： 神经元的传播系数系统稳定性之间存在联系

我们看上图，当传播系数大于 1 时，随着时间的发展，系统中信号的传播将会产生爆炸； 而当传播系数小于 1 时，由于每一次传播后激活神经元的个数都在变少，因此最终信号会呈指数消退； 只有当传播系数等于 1 时，系统才会保持相对的稳定。

我们将这种传播系数等于 1 的稳定状态称为临界状态，把传播系数大于 1 的情况称为超临界状态，小于 1 的情况称为亚临界状态。显然无论是亚临界还是超临界状态，都不利于信息的传递和处理。只有在临界状态，信息才能够通过神经元的活动把信息保持并传播下去。

Critical state and Reservoir Computing



Zeng et al. *Neural Networks 2019*

图 5：Critical state and Reservoir Computing

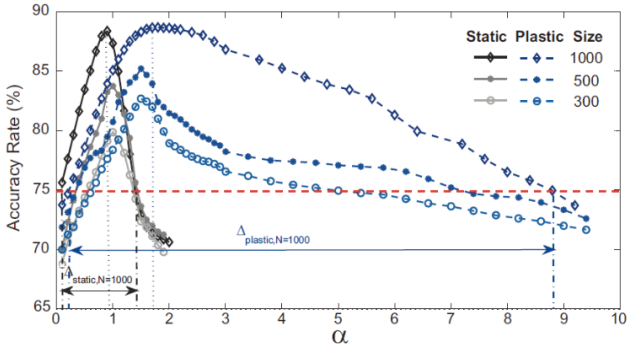
临界状态下，网络错误率往往会比较低。但从上图中可以看出，临界状态是非鲁棒的，稍微有一点扰动，其性能便会受到很大的影响。如何解决这一问题，使神经网络在保持高性能的情况下同时还具有较高的鲁棒性？大脑给了我们可以借鉴的答案：自适应机制。

Critical state and Reservoir Computing

Membrane potential: $\hat{h}_i(t)=\sum_{j=1}^{N} uJ_{ij}\delta(t-t_j^{SP}-\tau_D)+\delta_{i,\xi(t)}I^{ext}+I_i^{stf}(t)$

Short-term depression (STD): $\hat{J}_{ij}(t)=\frac{1}{\tau_J}(\frac{\alpha}{u}-J_{ij}(t))-uJ_{ij}(t)\delta(t-t_j^{SP})$

Levina et al. *Nat. Phys.* 2008



Zeng et al. *Neural Networks 2019*

图 6：自适应模型可以有效扩展临界状态宽度

神经科学家根据生物实验，提出了模拟模型，让网络模型能够自适应地学习传播系数。结果如上图所示，正常情况下，临界状态很窄；而采用自适应模型，临界状态的宽度便能大大增加。

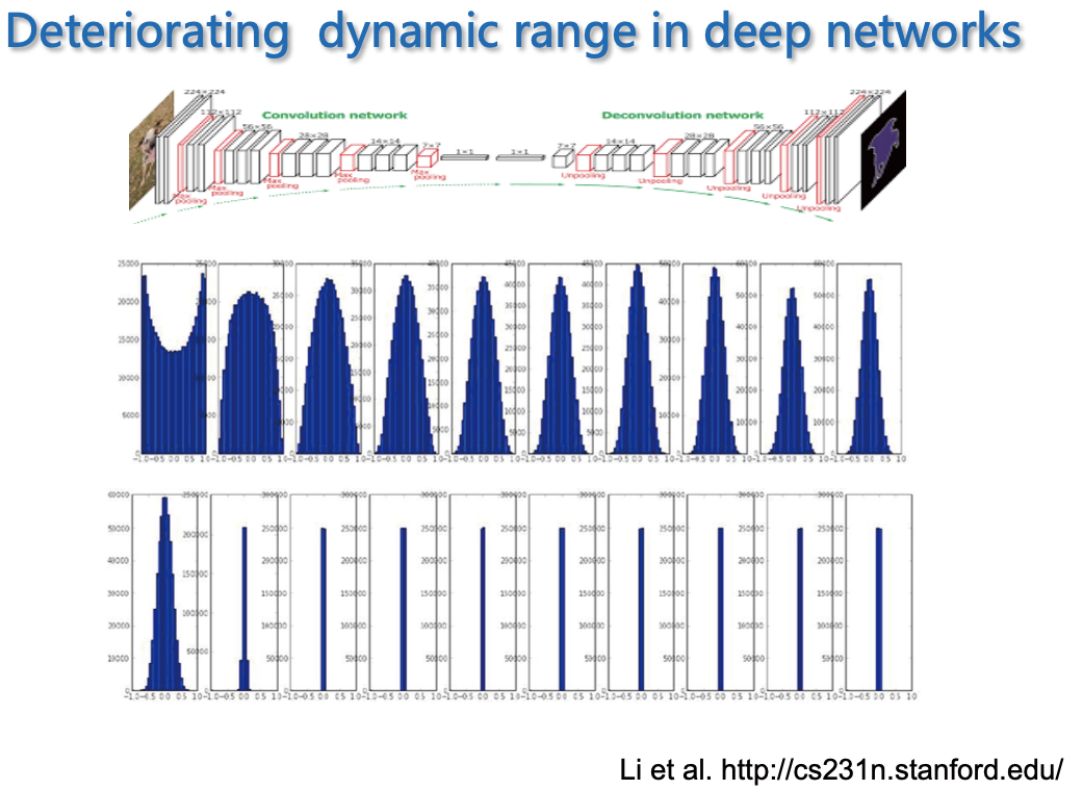


图 7：Deteriorating dynamic range in deep networks

同样的，在人工神经网络中，也存在着信息传输所带来的网络爆炸或快速消失问题，即所谓“梯度消散”。针对这一问题，目前主流的解决方案是用 Batch Norm 或 Layer Norm，也即在人工神经网络中，每两个处理层次之间添加一个专门的处理层；这个层的作用是把前面传来的信号做增强或衰减的调整，从而使后面层的反应不至太强或太弱。

但新增层必然会带来额外的计算负担。大脑就没有这种额外层。能否借鉴大脑的机制，通过某种方式，在不使用额外层的情况下，同时还能够保持网络的信息传播平衡呢?

三、计算层面：情境相关

相较于算法层面，在计算层面的借鉴稍显抽象。余山研究员结合他们近期发表在《Nature Machine Intelligence》上的一篇工作 (Continual Learning of Context-dependent Processing in Neural Networks)，做了相应的介绍。

人类作为智慧生物，最重要的特征便是能够 ” 适应环境变化，实现自身目的 ”。人类大脑不仅可以在新的环境中不断吸收新的知识，而且可以根据不同的环境灵活调整自己的行为。

作为对应，当前以 DNN 为代表的神经网络，尽管可以建立输入输出之间非常复杂的映射关系，用于识别、分类和预测。

但是一旦学习阶段结束，它所能做的操作就固化了，既难以方便的学习新的映射，也不能对实际环境中存在情境信息 (比如自身状态、环境变化、任务变化等) 做出灵活的响应，难以满足复杂多变的需求，即缺少情境依赖学习 (contextual–dependent learning) 的能力。那么，我们如何借鉴脑科学知识呢?

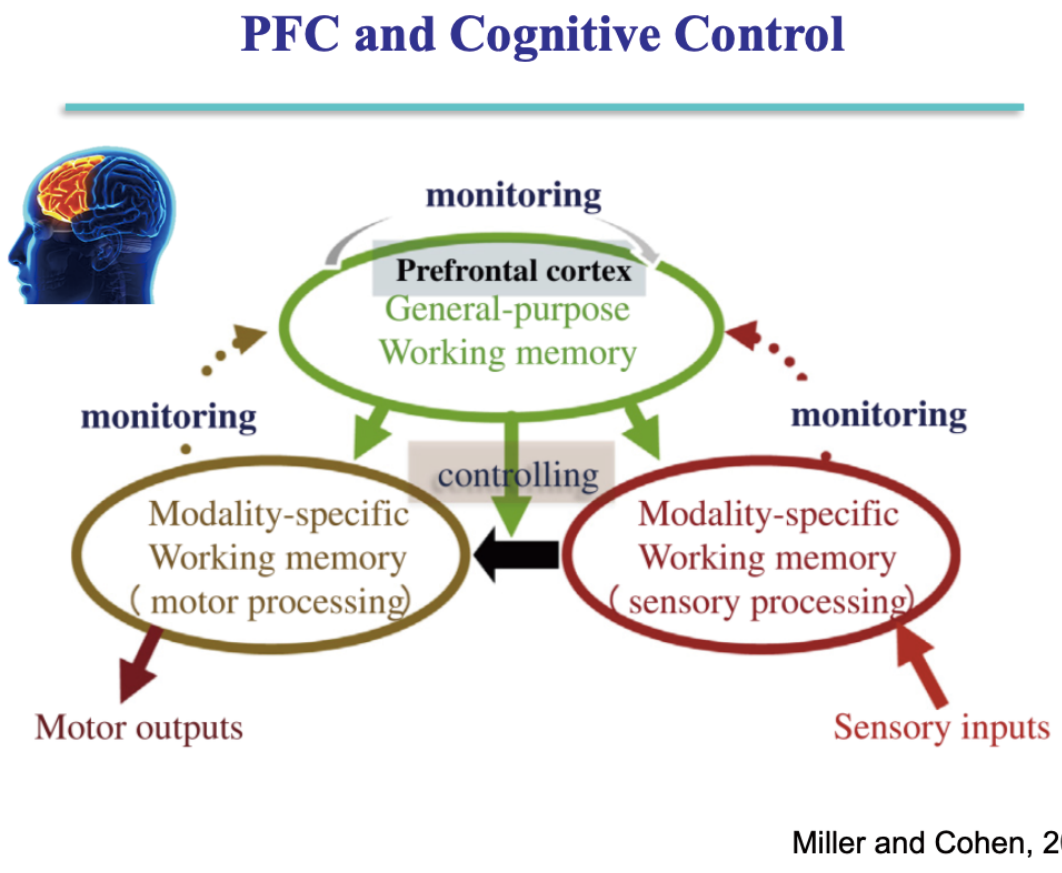
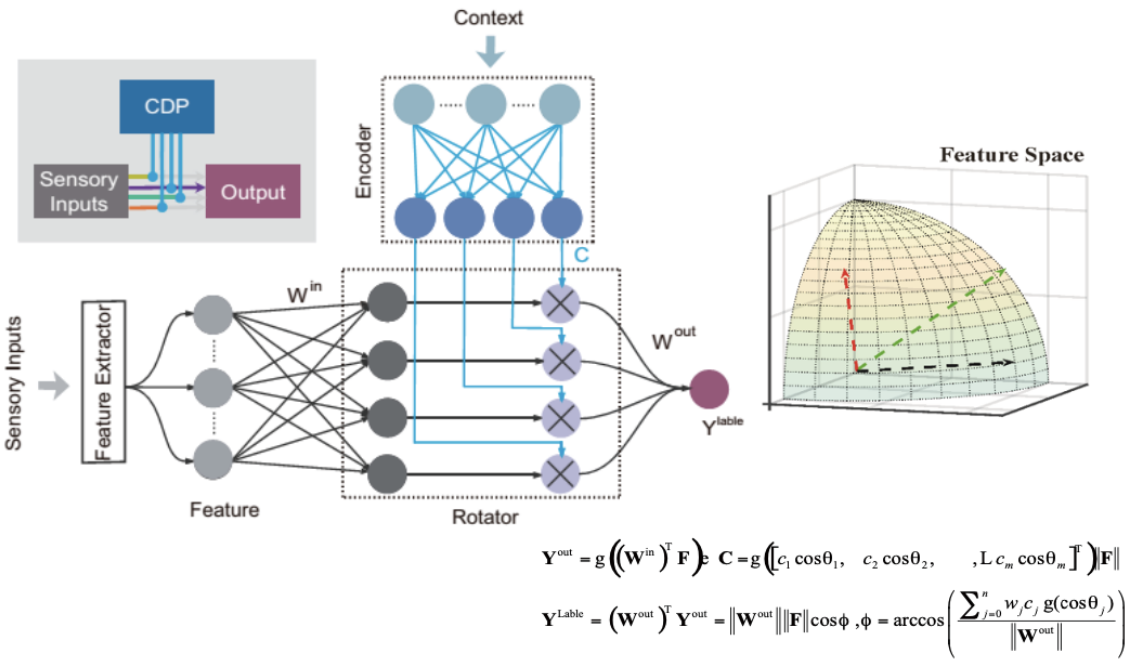


图 8：PFC and Cognitive Control

据脑科学家的研究表明，大脑的结构，除了感觉输入、运动输出这个通路之外，还存在一个调控的通路 (主要在大脑前额叶发挥作用，因此也可以说，前额叶区决定了人的随机应变能力)。这个调控通路在很大程度上决定了人的灵活应变能力。

PFC-like module



Zeng G*, Chen Y*, Cui B and Yu S. *Nature Machine Intelligence* 2019

图 9：PFC–like module

受此启发，余山等人提到了一种 PFC–like 的新网络架构，在输入输出之间加入了一个情境处理模块 (CDP)。CDP 模块的作用便是在输入输出之间，根据 Context 对结果进行旋转，从而能够依据上下文动态调整网络内部信息。

它包括两个子模块：1. 编码子模块，其负责将情境信息编码为适当的控制信号；2.“旋转”子模块，其利用编码模块的控制信号处理任务输入 (由于其功能上相当于将特征向量在高维空间上进行了旋转，故称为“旋转”子模块)。结果喜人！

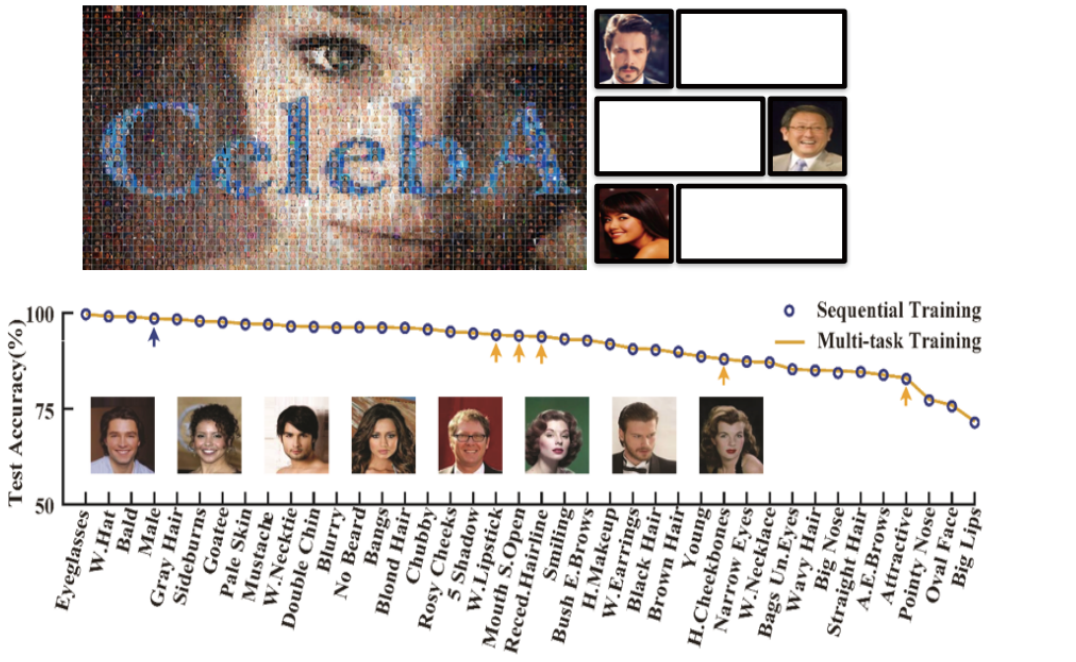


图 10：Context-dependent face recognition（注：同一个分类器对于同样的输入，连续学习 40 种不同人脸属性的分类任务，正确率与用 40 个分类器的系统几乎一致。）

他们在 CelebA 数据集上进行测试。按照传统的模型，针对数据集上的 40 个类型，需要训练 40 个模型才能完成任务，而采用 CDP 模块后，一个模型能解决所有分类问题，且性能不降。若想进一步了解这个奇妙的思想，可参看文章：《国内首发 Nature 子刊 Machine Intelligence 论文：思想精妙，或对 DNN 有重大改进！》

四、学习层面：连续学习和情境依赖

学习层面，神经网络面临的一个重要问题是灾难性遗忘，即神经网络在学习不同的任务时，如果不是把不同任务的训练样本混在一起去训练，往往在学习新的任务时候，网络就会把从旧任务中学到的知识忘掉。

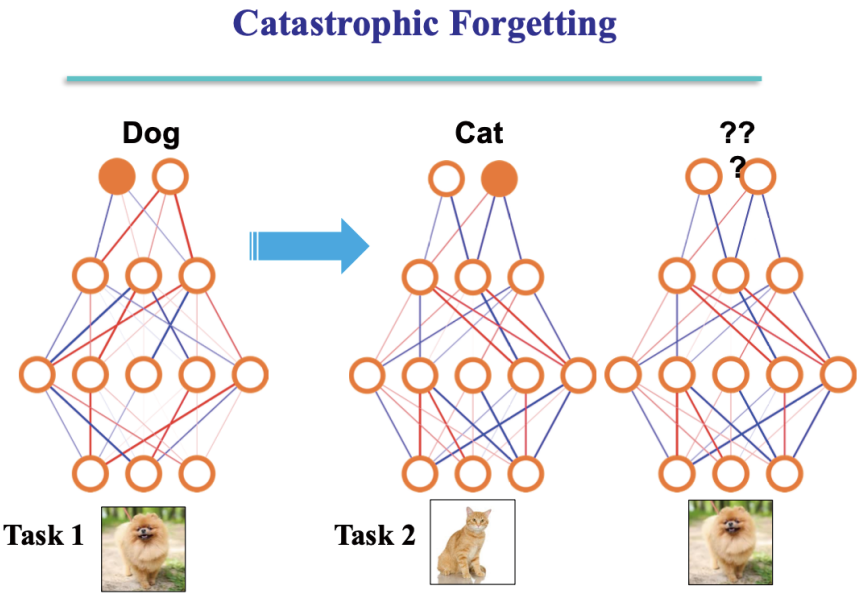


图 11：Catastrophic forgetting

以上图为例，先让神经网络识别「狗」，得到一个性能非常高的网络；继而再让网络去学习识别「猫」，这时网络的权重就会重新调整；学完之后再拿来去识别「狗」，神经网络的性能就会大幅下降，甚至不能使用。

原因就在于，当学习「猫」的任务时，网络把针对「狗」的任务学到的知识给忘了。然而，人脑却没有这种所谓「灾难遗忘」的问题。人类先后顺序地学习不同的任务，最后识别能力还能不断提升。针对这一问题，余山研究员在上面提到的那篇文章中提出一种称为「正交权重修改 (Orthogonal Weights Modification，OWM) 」的算法。

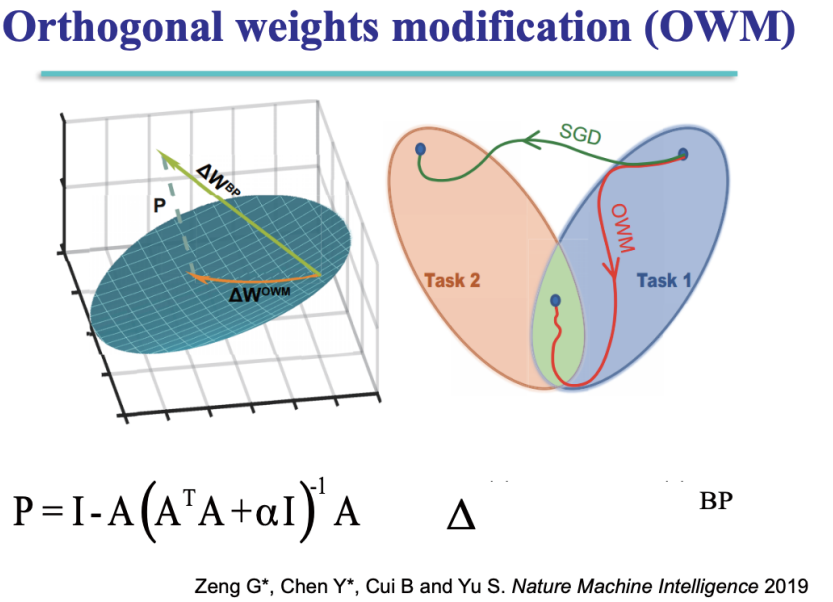
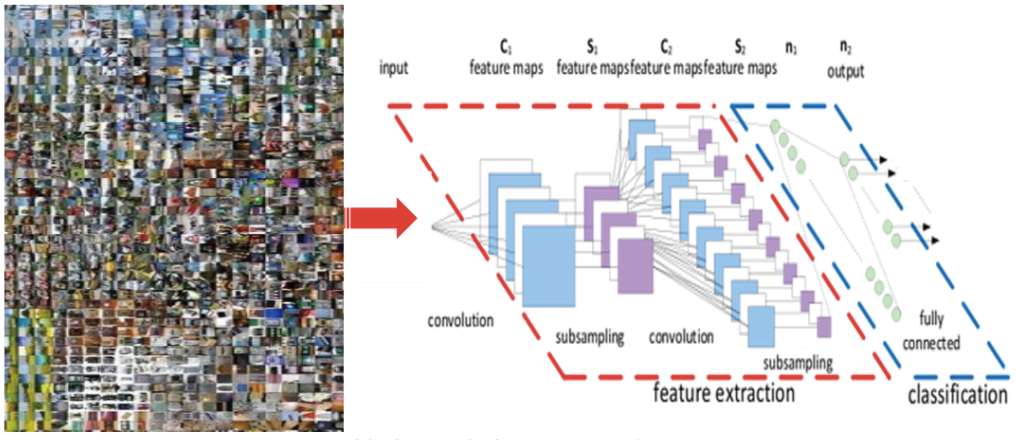


图 12：OWM 算法原理示意图。(a): 在权重更新时，OWM 算法只保留传统 BP 算法计算的权重增量中与历史任务输入空间正交的部分；(b): 在新任务中，OWM 算法将神经网络对解的搜索范围约束在旧任务的解空间中。

OWM 算法的核心思想很简单，即通过 P 映射之后，学习新任务的解仍然在旧任务的解空间当中。正如其名“正交权重修改”，在学习新任务时，只在旧任务输入空间正交的方向上修改神经网络权重。如此，权重增量几乎不与以往任务的输入发生作用，从而保证了网络在新任务训练过程中搜索到的解，仍处在以往任务的解空间中。数学上，OWM 通过正交投影算子 P 与误差反传算法得到的权重增量 Δ w 作用来实现其目的，即最终的权重增量Δw=kpΔw，这里 k 为系数。OWM 算法实现了对网络中已有知识的有效保护，并可以与现有梯度反传算法完全兼容。

Performance on larger dataset --- ImageNet

Data Set	Classes	Feature Extractor	Concurrent Training by SGD (%)	Sequential Training by OWM (%)	Sequential Training by SGD (%)
Image Net	1000	ResNet152	78.31	75.24	4.27

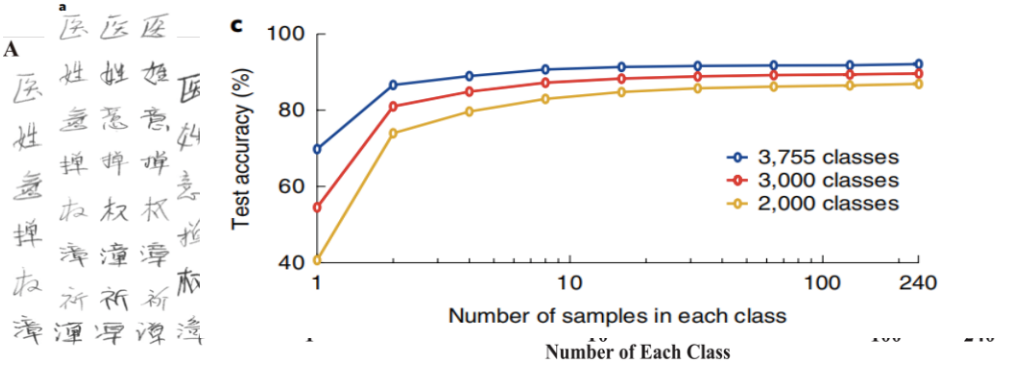


Zeng G*, Chen Y*, Cui B and Yu S. *Nature Machine Intelligence* 2019

图 13：Performance on larger dataset---ImageNet

如上图所示，余山等人在 ImageNet 上做了测试，选取 1000 个类，特征提取器使用 ResNet152，在训练分类器时:

1) 当采用传统的 SGD 方法，任务混合训练时，准确率为 78.31%；
2) 在采用 SGD，但所有任务顺序训练时，准确率直降到 4.27%，这正是「灾难性遗忘」的结果；
3) 当采用 OWM 方法，任务顺序训练时，结合经过预训练的特征提取器，准确率能够达到 75.24%，性能媲美于 SGD 的混合训练。



Data Set	Classes	Feature Extractor	Concurrent Training by SGD (%)	Sequential Training by OWM (%)	Sequential Training by SGD (%)
CASIA-HWDB1.1	3755	ResNet18	97.46	93.46	35.86

C L. CASIA Online and Offline Chinese Handwriting Databases[J]. NLPR, 2011.

Zeng G*, Chen Y*, Cui B and Yu S. *Nature Machine Intelligence* 2019

图 14：Performance on HWDB dataset

余山等人同样在手写字数据集 HWDB 上进行了测试，包含 3755 个类，特征提取器选用 ResNet18，同样可以看到，采用 OWM 顺序训练分类器依然能够保持较高的性能。

值得一提的是，算法具有优良的小样本学习能力，以手写体汉字识别为例，基于预训练的特征提取器，系统可以从仅仅数个正样本中就能连续的学习新的汉字。上图中显示在 3755 个类（汉字）上，仅需要在 10 个类上进行连续学习，便能够达到 90% 以上的性能。

OWM 算法有效地克服了灾难性遗忘的难题，使得单个神经网络不仅可以先学「狗」再学「猫」，而且可以逐渐的学习多达数千个类型的识别。这一新型学习算法和前面介绍的情境依赖处理（CDP）模块配合，能够使人工神经网络具备强大的连续学习和情境依赖学习能力。

其中，OWM 算法可以有效克服神经网络中的灾难性遗忘，实现连续学习； 而受大脑前额叶皮层启发的 CDP 模块可以有效整合情境信息，调制神经网络的信息处理过程。二者结合便有望让智能体通过连续不断的学习去适应复杂多变的环境，从而逐步逼近更高水平的智能。

五、先验知识、语义理解和记忆

除了上面四个层次的借鉴之外，余山老师还介绍了如何将先验知识压缩并注入神经网络、从符号计算到语义理解、从有监督的分类训练到无监督的重构和预测等类脑计算的思路。如何将先验知识压缩并注入神经网络。

认知学家曾经做过一个实验，即从小教一个黑猩猩学习语言，发现黑猩猩在语言学习上远远不能达到人类的高

度。这说明我们人类大脑有先天的神经结构能够让我们容易学习语言，这种先天结构即为先验知识。作为对比，当前的神经网络基本上没有先验知识，都得从头学起。

那么我们是否可以借鉴大脑积累先验知识的机制，来设计人工神经网络呢？从符号计算到语义理解。目前的自然语言处理系统训练的材料是语料，纯粹是文字或符号。

以中文屋 (Chinese Room) 实验为例，里面纯粹是做一些非常简单的信息处理工作，只是一个符号到符号的处理过程，并没有真正理解内在的含义。因此 NLP 的研究，若想克服这个问题，未来必然需要向大脑学习。有监督的分类训练到无监督的重构和预测。

当前，训练好的做分类任务的神经网络在复杂环境下往往性能并不好。但对比一下，人类的视觉系统并没有使用监督信号去训练分类任务，例如小孩学习识别物体，完全是靠自监督的方式看这个世界的。因此，真正的强人工智能可能并不是现在这种端到端的有监督训练，而是采用类脑的分阶段的、包含无监督或自监督的训练方式。最近机器学习领域的进展，也说明了这一策略正逐渐受到人们的关注。

六、结语

余山研究员总结道，虽然我们对于大脑的了解尚不完备，生物脑和人工神经网络的结构也有很大的差异，但是这并不是开展类脑计算研究的本质障碍。

神经科学和认知科学的研究已经发现了大脑的很多机制性原理，这些知识足够指导我们不断改善智能系统的设计，最终有望实现在不同层面上受脑启发的更加强大和高效的人工智能系统。