机器与人类视觉能力的差距（2）

> 本文属于个人观点，跟本人在职公司的立场无关。由于最近 GitHub 服务器在国内访问速度严重变慢，虽然经过大幅度压缩尺寸，文中的图片仍然可能需要比较长时间才能加载。这篇文章揭示了 AI 领域重要的谬误和不实宣传，为了阻止愚昧的蔓延，我鼓励大家转发这篇文章和它的后续，转发时只需要注明作者和出处就行。

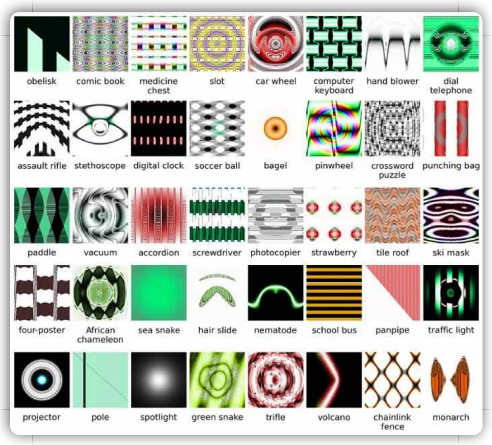
这是这个系列文章的第二集，在这一集中，我想详细分析一下 AI 领域到底理解多少人类神经系统的构造。

神经网络为什么容易被欺骗

“神经网络”与人类神经系统的关系是是很肤浅的。等你理解了所谓“神经网络”，就会明白它跟神经系统几乎没有一点关系。“神经网络”只是一个误导性质的 marketing 名词，它出现的目的只是为了让外行产生不明觉厉的效果，以为它跟人类神经系统有相似之处，从而对所谓的“人工智能”信以为真。

其实所谓“神经网络”应该被叫做“可求导编程”。说穿了，所谓“神经网络”，“机器学习”，“深度学习”，就是利用微积分，梯度下降法，用大量数据拟合出一个函数，所以它只能做拟合函数能做的那些事情。

用了千万张图片和几个星期的计算，拟合出来的函数也不是那么可靠。人们已经发现用一些办法生成奇怪的图片，能让最先进的深度神经网络输出完全错误的结果。



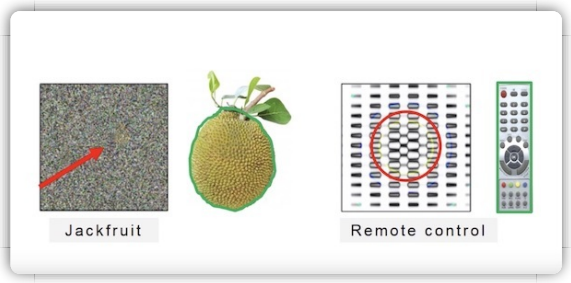
（图片来源：http://www.evolvingai.org/fooling）

神经网络为什么会有这种缺陷呢？因为它只是拟合了一个“像素=>名字”的函数。这函数碰巧能区分训练集里的图片，却不能抓

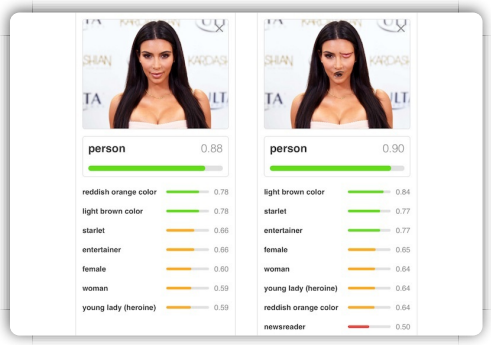
住物体的结构和本质。它只是像素级别的拟合，所以这里面有很多空子可以钻。

深度神经网络经常因为一些像素，颜色，纹理匹配了物体的一部分，就认为图片上有这个物体。它无法像人类一样理解物体的结构和拓扑关系，所以才会被像素级别的肤浅假象所欺骗。

比如下面两个奇怪的图片，被认为是一个菠萝蜜和一个遥控器，仅仅因为它们中间出现了相似的纹理。



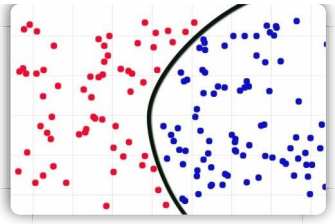
另外，神经网络还无法区分位置关系，所以它会把一些位置错乱的图片也识别成某种物体。比如下面这个，被认为是一张人脸，却没发现五官都错位了。



神经网络为什么会犯这种错误呢？因为它的目标只是把训练集里的图片正确分类，提高“识别率”。至于怎么分类，它可以是毫无原则的，它完全不理解物体的结构。它并没有看到“叶子”，“果皮”，“方盒子”，“按钮”，它看到的只是一堆像素纹理。因为训练集里面的图片，出现了类似纹理的都被标记为“菠萝蜜”和“遥控器”，没有出现这纹理的都被标记为其它物品。所以神经网络找到了区分它们的“分界点”，认为看到这样的纹理，就一定是菠萝蜜和遥控器。

我试图从神经网络的本质，从统计学来解释这个问题。神经网络其实是拟合一个函数，试图把标签不同的样本分开。拟合出来的函数试图接近一个“真实分界线”。所谓“真实分界线”，是一个完全不会错的函数，也就是“现实”。

数据量小的时候，函数特别粗糙。数据量大了，就逐渐逼近真实分界线。但不管数据量如何大，它都不可能得到完全准确的“解析解”，不可能正好抓住“现实”。



除非现实函数特别简单，运气特别好，否则用数据拟合出来的函数，都会有很多小“缝隙”。以上的像素攻击方法，就是找到真实分界线附近，“缝隙”里面的样本，它们正好让拟合函数出现分类错误。

人的视觉系统是完全不同的，人直接就看到了事物是什么，看到了“解析解”，看到了“现实”，而没有那个用数据逼近的过程，所以除非他累得头脑发麻或者喝了酒，你几乎不可能让他判断错误。

退一步来看，图像识别所谓的“正确分类”都是人定义的。是人给了那些东西名字，是许多人一起标注了训练用的图片。所以这里所谓的“解析解”，“现实”，全都是人定义的。一定是某人看到了某个事物，他理解了它的结构和性质，然后给了它一个名字。所以别的人也可以通过理解同一个事物的结构，来知道它是什么。

神经网络不能看到事物的结构，所以它们也就难以得到精确的分类，所以机器在图像识别方面是几乎不可能超越人类的。现在所谓的“超人类视觉”的深度学习模型，大部分都是欺骗和愚弄大众。使用没有普遍性的数据集，使用不公平的准确率标准来对比，所以才显得机器好像比人还厉害了。这是一个严重的问题，在后面我会详细分析。

神经网络训练很像应试教育

神经网络就像应试教育训练出来的学生，他们的目标函数是“考高分”，为此他们不择手段。等毕业工作遇到现实的问题，他们就傻眼了，发现自己没学会什么东西。因为他们学习的时候只是在训练自己“从 ABCD 里区分出正确答案”。等到现实中没有 ABCD 的时候，他们就不知道怎么办了。

深度学习训练出来的那些“参数”是不可解释的，因为它们存在的目的只是把数据拟合出来，把不同种类的图片分离开，而没有什么意义。AI 人士喜欢给这种“不可解释性”找借口，甚至有人说：“神经网络学到的数据虽然不可解释，但它却出人意料的有效。这些学习得到的模型参数，其实就是知识！”

这些模型真的那么有效吗？那为什么能够被如此离谱的图片所欺骗呢？说“那就是知识”，这说法简直荒谬至极，严重玷污了“知

识”这个词的意义。这些“学习”得到的参数根本就不是本质的东西，不是知识，真的就是一堆毫无道理可言的数字，只为了降低“误差”，能够把特征空间的图片区分开来，所以神经网络才能被这样钻空子。

说这些参数是知识，就像在说考试猜答案的技巧是知识一样可笑。“另外几套题的第十题都是 B，所以这套题的第十题也选 B”…… 深度学习拟合函数，就像拿历年高考题和它们的答案来拟合函数一样，想要不上课，不理解科目知识就做出答案来。有些时候它确实可以蒙对答案，但遇到前所未见的题目，或者题目被换了一下顺序，就傻眼了。

人为什么可以不受这种欺骗呢？因为人提取了高级的拓扑结构，不是瞎蒙的，所以人的判断不受像素的影响。因为提取了结构信息，人的观察是具有可解释性的。如果你问一个小孩，为什么你说这是一只猫而不是一只狗呢？她会告诉你：“因为它的耳朵是这样的，它的牙是那样的，它走路的姿势是那样的，它常常磨爪子，它用舌头舔自己……”

做个实验好了，你可以问问你家孩子这是猫还是狗。如果是猫，为什么他们认为这是一只猫而不是一只狗？

神经网络看到一堆像素，很多层处理之后也不知道是什么结构，分不清“眼睛”，“耳朵”和“嘴”，更不要说“走路”之类的动态概念了，所以它也就无法告诉你它认为这是猫的原因了。拟合的函数碰巧把这归成了猫，如果你要追究原因，很可能是肤浅的：图片上有一块像素匹配了图片库里某只猫的毛色纹理。

有一些研究者把深度神经网络的各层参数拆出来，找到它们对应的图片中的像素和纹理，以此来证明神经网络里的参数是有意义的。咋一看好像有点道理，原来“学习”就能得到这么多好像设计过的滤镜啊！可是仔细一看，里面其实没有多少有意义的内容，因为它们学到的参数只是能把那些图片类别分离开。

所以人的视觉系统很可能是跟深度神经网络原理完全不同的，或者只有最低级的部分有相似之处。

“神经网络”与人类神经元的关系是肤浅的

为什么 AI 人士总是认为视觉系统的高级功能都能通过“学习”得到呢？非常可能的事情是，人和动物视觉系统的“结构理解”，“3D建模”功能不是学来的，而是早就固化在基因里了。想一想你生下来之后，有任何时候看到世界是平面的，毫无关联的像素吗？

所以我觉得，人和动物生下来就跟现有的机器不一样，结构理解所需的硬件在胚胎里就已经有了，只等发育和激活。人是有学习能力，可是人的学习是建立在结构理解之上，而不是无结构的像素。另外人的“学习”很可能处于比较高的层面，而不是神经元那么“底层”的。人的神经系统里面并没有机器学习那种 back-propagation。

纵使你有再多的数据，再多的计算力，你能超越为期几十亿年的，地球规模的自然进化和选择吗？与其自己去“训练”或者“学习”，不如直接从人身上抄过来！但问题是，我们真的知道人的视觉系统是如何工作的吗？

神经科学家们其实并没有完全搞明白人类视觉系统是如何工作的。就像所有的生物学领域一样，人们的理解仍然是很粗浅的。神经网络与人类视觉系统的关系是肤浅的。每当你质疑神经网络与人类视觉系统的关系，AI 研究者就会抬出 Hubel & Wiesel 在 1959 年拿猫做的那个实验：“有人已经证明了人类视觉系统就是那样工作的！” 如此的自信，不容置疑的样子。

我问你啊，如果我们在 1959 年就已经知道人类视觉系统的工作原理细节，为什么现在还各种模型改来改去，训练来训练去呢？直接模仿过来不就行了？所以这些人的说法是自相矛盾的。

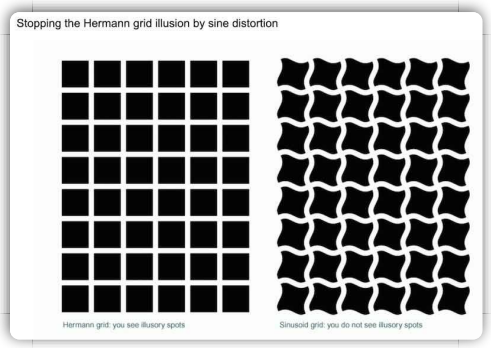
你想过没有，为什么到了 2019 年，AI 人士还拿一个 60 年前的实验来说明问题？这 60 年来就没有新的发现了吗？而且从 H&W 的实验你可以看出来，它只说明了猫的视觉神经有什么样的底层功能（能够做“线检测”），却没有说那就是全部的构造，没说上层的功能都是那样够构造的。

H&W 的实验只发现了最底层的“线检测”，却没有揭示这些底层神经元的信号到了上层是如何组合在一起的。“线检测”是图像处理的基础操作。一个能够识别拓扑结构的动物视觉系统，理所当然应该能做“线检测”，但它应该不止有这种低级功能。

视觉系统应该还有更高级的结构，H&W 的实验并没能回答这个问题，它仍然是一个黑盒子。AI 研究者们却拿着 H&W 的结果大做文章，自信满满的声称已经破解了动物视觉系统的一切奥秘。

那些说“我们已经完全搞明白了人类视觉是如何工作”的 AI 人士，应该来看看这个 2005 年的分析 Herman grid 幻觉现象的

幻灯片。这些研究来自 Schiller Lab，MIT 的脑科学和认知科学实验室。通过一系列对 Herman grid 幻觉图案的改动实验，他们发现长久以来（从 1960 年代开始）对产生这种现象的理解是错误的：那些暗点不是来自视网膜的“边沿强化”功能。他们猜想，这是来自大脑的 V1 视觉皮层的 S1 “方向选择”细胞。接着，另一篇 2008 年的 paper 又说，Schiller 的结果是不对的，这种幻觉跟那些线条是直的有关系，因为你如果把那些白线弄弯，幻觉就消失了。然后他们提出了他们自己的，新的“猜想”。



从这种研究的方式我们可以看出，即使是 MIT 这样高级的研究所，对视觉系统的研究还处于“猜”的阶段，把人脑作为黑盒子，拿一些图片来做“行为”级别的实验。他们并没有完全破解视觉系统，看到它的“线路”和“算法”具体如何工作，而是给它一些输入，测试它的输出。这就是“黑盒子”实验法。以至于很多关于人类视觉的理论都不是切实而确定的，很可能是错误的猜想。

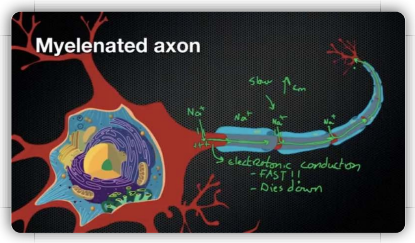
脑科学发展到今天也还是如此，AI 领域相对于脑科学的研究方式，又要低一个级别。2019 年了，仍然抬出神经科学家 1959 年的结果来说事。闭门造车，对人家的最新成果一点都不关心。现在的深度神经网络模型基本是瞎蒙出来的。把一堆像素操作叠在一起，然后对大量数据进行“训练”，以为这样就能得到所有的视觉功能。

动物视觉系统里面真有“反向传导”（back-propagation）这东西吗？H&W 的实验里面并没有发现 back-propagation。实际上神经科学家们至今也没有发现神经系统里面有 back-propagation，因为神经元的信号传递机制不能进行“反向”的通信。很多神经科学家的结论是，人脑里面进行 back-propagation 不大可能。

所以神经网络的各种做法恐怕没有受到 H&W 实验的多大启发。只是靠这么一个肤浅的相似之处来显得自己接近了“人类神经系统”。现在的所谓“神经网络”，其实只是一个普通的数学函数的表达式，里面唯一起作用的东西其实是微积分，所谓 back-propagation，就是微积分的求导操作。神经网络的“训练”，就是反复求导数，用梯度下降方法进行误差最小化，拟合一个函

数。这一切都跟神经元的工作原理没什么关系，完全就是数学。

为了消除无知带来的困惑，你可以像我一样，自己去了解一下人类神经系统的工作原理。我推荐你看看这个叫《Interactive Biology》的 YouTube 视频系列。你可以从中轻松地理解人类神经系统一些细节：神经元的工作原理，视觉系统的原理，眼睛，视网膜的结构，听觉系统的工作原理，等等。神经学家们对此研究到了如此细节的地步，神经传导信息过程的每一个细节都展示了出来。



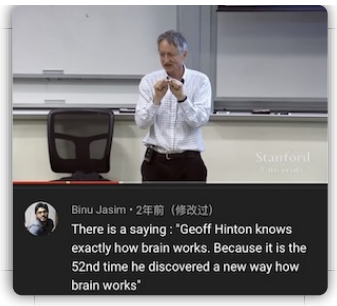
AI 研究者并不知道人脑如何工作

AI 领域真的理解人脑如何工作吗？你可以参考一下这个演讲："Can the brain do back-propagation?"（人脑能做 back-propagation 吗？）。演讲人是深度学习的鼻祖级人物 Geoffrey Hinton。他和其它两位研究者（Yoshua Bengio 和 Yann LeCun），因为对深度学习做出的贡献，获得了 2018 年的图灵奖。演讲一开头 Hinton 说，神经科学家们说人脑做 back-propagation 是不可能的，然后他开始证明这是可能的，依据神经元的工作原理，back-propagation 如何能用人脑神经元来实现。

是的，如果你有能力让人脑按你的"算法"工作的话，神经元组成的系统也许真能做 back-propagation，可是人脑是你设计的吗？很可惜我们无法改变人脑，而只能去"发现"它到底是如何工作。这不是人脑"能不能"的问题，而是"做不做"的问题。研究人脑是一个科学发现工作，而不是一个工程设计工作。

看了这个演讲，我觉得 AI 人士已经进入了一种"上了天"的状态。他们坚定的认为自己的模型（所谓的"神经网络"）就是终极答案，甚至试图把人脑也塞进这个模型，设想人脑神经元如何能实现他们所谓的"神经网络"。可是他们没有发现，人脑的方式也许比他们的做法巧妙很多，根本跟他们的"神经网络"不一样。

从这个视频我们也可以看出，神经科学界并不支持 AI 领域的说法。AI 领域是自己在那里瞎猜。视频下面有一条评论我很欣赏，他用讽刺的口气说："Geoff Hinton 确切地知道人脑是如何工作的，因为这是他第 52 次发现人脑工作的新方式。"



AI 人的盲目信仰

AI 人士似乎总是有一种不切实际的“信仰”或者“信念”，他们坚信机器一定可以具有人类一样的智能，总有一天能够在所有方面战胜人类。总是显示出一副“人类没什么了不起”的心态，张口闭口拿“人类”说事，好像他们自己是另外一个物种，已经知道人类的一切能力，有资格评判所有人的智力似的。

我不知道是什么导致了这种“AI 宗教”。有句话说得好：“我所有的自负都来自我的自卑，所有的英雄气概都来自于我内心的软弱，所有的振振有词都因为心中满是怀疑。” 似乎是某种隐藏很深的自卑和怨恨，导致了他们如此的坚定和自负。一定要搞出个超越所有人的机器才善罢甘休，却没发现人类智能的博大精深已经从日常生活的各种不起眼的小事透露出来。

他们似乎看不到世界上有各种各样，五花八门的人类活动，每一种都显示出奇迹般的智能。连端茶倒水这么简单的事情，都包含了机器望尘莫及的智能，更不要说各种体育运动，音乐演奏，各种研究和创造活动了。就连比人类“低级”一点的动物，各种宠物，家畜家禽，飞鸟走兽，甚至昆虫，全都显示出足以让人敬畏的智能。他们对所有这些奇迹般的事物视而不见，不是去欣赏他们的精巧设计和卓越表现，而是坐井观天，念叨着“机器一定会超越人类”。

他们似乎已经像科幻电影似的把机器当成了一个物种，像是保护“弱势群体”一样，要维护机器的“权益”和“尊严”。他们不允许其他人质疑这些机器，不允许你说它们恐怕没法实现人类一样的智能。总之机器在他们心理已经不再是工具，而是活的生命，甚至是比人还高级的生命。

对此你可以参考另一个 Geoffrey Hinton 的采访视频，录制于今年 5 月份的 Google 开发者大会（Google I/O ‘19）。

从这个视频里面我看到了许多 AI 人士盲目信仰和各种没有根据的说法的来源，因为这些说法全都集中而强烈的体现在了 Hinton 的谈话中。他如此的坚信一些没有根据的说法，不容置疑地把它们像真理一样说出来，却没有任何证据。有时候主持人

都不得不采用了有点怀疑的语气。

Hinton 在采访中有以下说法：

1. “神经网络被设计为像人脑的工作原理。”
2. “等神经网络能够跟人对话，我们就能用它来进行教育工作了。”
3. “神经网络终究会在所有事情上战胜人类。”
4. “我们不都是神经网络吗？”（先后强调了两次）
5. “..... 所以神经网络能够实现人类智能的一切功能。这包括感情，意识等。”
6. “人们曾经认为生命是一种特殊的力量，现在生物学解释了生命的一切。人们现在仍然认为意识是特殊的，可是神经网络将会说明，意识并没有什么特别。”

他的这些说法都是不准确，不科学，没有根据的。

我发现每当主持人用稍微怀疑的语气问：“这真的可以实现吗？” Hinton 就会回答：“当然能。我们不都是神经网络吗？” 这里有一个严重的问题，那就是他所谓的“神经网络”，其实并不是人脑里面的神经元连成的网络。AI 领域的“神经网络”只是他们自己的数学模型，是他们自己给它起名叫“神经网络”而已。所以他的这种“证明”其实是在玩文字游戏：“因为我们都是神经网络，所以神经网络能够实现一切人类智能，感情，甚至意识本身！”

前面的“神经网络”和后面的“神经网络”完全是两回事。我们是“神经网络”吗？我们的脑子里是有神经元，神经元貌似连成了一个网络，可是它的结构却跟 AI 领域所谓的“神经网络”是两回事，工作原理也非常不一样。Hinton 面对问题作出这样的回答，是非常不科学，不负责任的。

最后关于生命，感情和意识的说法，我也很不认同。虽然生物学解释了生命体的各种构造和原理，可是人们为什么仍然没能从无生命的物质制造出有生命的事物呢？虽然人们懂得那么多生物学，生物化学，有机化学，甚至能合成出各种蛋白质，可是为什么没能把这些东西组装在一起，让它“活”起来呢？这就像你能造

出一些机器零件，可是组装起来之后，发现这机器不转。你不觉得是因为少了点什么吗？生物学发展了这么久，我们连一个最简单的，可以说是“活”的东西都没造出来过，你还能说“生命没什么特别的”吗？

这说明生物学家们虽然知道生命体的一些工作原理，却没有从根本上搞明白生命到底是什么。也就是说人们解决了一部分“how”问题（生命体如何工作），却不理解“what”和“why”（生命是什么，为什么会出现生命）。

实际上生物学对生命体如何工作（how）的理解都还远远不够彻底，这就是为什么我们还有那么多病无法医治，甚至连一些小毛病都无法准确的根治，一直拖着，只是不会马上致命而已。“生命是什么”的 what 问题仍然是一个未解之谜，而不像 Hinton 说的，全都搞明白了，没什么特别的。

也许生命就是一种特别的东西呢？也许只有从有生命的事物，才能产生有生命的事物呢？也许生命就是从外星球来的，也许就是由某种更高级的智慧设计出来的呢？这些都是有可能的。真正的科学家应该保持开放的心态，不应该有类似“人定胜天”这样的信仰。我们的一切结论都应该有证据，如果没有我们就不应该说“一定”或者“必然”，说得好像所有秘密全都解开了一样。

对于智能和意识，我也是一样的态度。在我们没有从普通的物质制造出真正的智能和意识之前，不应该妄言理解了关于它们的一切。生命，智能和意识，比有些人想象的要奇妙得多。想要“人造”出这些东西，比 AI 人士的说法要困难许多。

有心人仔细观察一下身边的小孩子，小动物，甚至观察一下自己，就会发现它们的“设计”是如此的精巧，简直不像是随机进化出来的，而是由某个伟大的设计者创造的。46 亿年的时间，真的够进化和自然选择出这样聪明的事物吗？

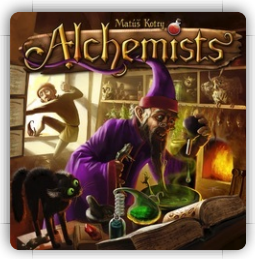
别误会了，我是不信宗教的。我觉得宗教的圣经都是小人书，都是某些人吓编的。可是如果你坚定的相信人类和动物的这些精巧的结构都是“进化”来的，你坚定的相信它们不是什么更高级的智慧创造出来的，那不也是另外一种宗教吗？你没有证据。没有证据的东西都只是猜想，而不能坚信。

好像扯远了……

总之，深度学习的鼻祖级人物说出这样多信念性质的，没有根据的话，由此可见这个领域有多么混沌。另外你还可以从他的谈话中看出，他所谓的“AI”都是各种相对容易的识别问题（语音识别，图像识别）。他并没有看清楚机器要想达成“理解”有多困难。而“识别”与“理解”的区别，就是我的这篇文章想澄清的问

题。

炼丹师的工作方式



设计神经网络的“算法工程师”，“数据科学家”，他们工作性质其实很像“炼丹师”（alchemist）。拿个模型这改改那改改，拿海量的图片来训练，“准确率”提高了，就发 paper。至于为什么效果会好一些，其中揭示了什么原理，模型里的某个节点是用来达到什么效果的，如果没有它会不会其实也行？不知道，不理解。甚至很多 paper 里的结果无法被别的研究者复现，存在作假的可能性。

我很怀疑这样的研究方式能够带来什么质的突破，这不是科学的方法。如果你跟我一样，把神经网络看成是用“可求导编程语言”写出来的代码，那么现在这种设计模型的方法就很像“一百万只猴子敲键盘”，总有一只能敲出“Hello World！”

许多数学家和统计学家都不认同 AI 领域的研究方式，对里面的很多做法表示不解和怀疑。为此斯坦福大学的统计学系还专门开了一堂课 Stats 385，专门讨论这个问题。课堂上请来了一些老一辈的数学家，一起来分析深度学习模型里面的各种操作是用来达到什么目的。有一些操作很容易理解，可是另外一些没人知道是怎么回事，这些数学家都看不明白，连设计这些模型的炼丹师们自己都不明白。

所以你也许看到了，AI 研究者并没能理解人类视觉系统的工作原理，许多的机器视觉研究都是在瞎猜。在接下来的续集中，我们会看到他们所谓的“超人类识别率”是如何来的。

请看下一篇：机器与人类视觉能力的差距（3）