机器与人类视觉能力的差距（3）

> 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三集，在这一集中，我想讲讲 AI 领域所谓的“超人类识别率”是怎么来的，以及由于对机器视觉的盲目信任所导致的灾难性后果。

“超人类准确率”的迷雾

我发现神经网络在测试数据的可靠性，准确率的计算方法上，都有严重的问题。

神经网络进行图像识别，所谓“准确率”并不是通过实际数据测出来的，而是早就存在那里的，专用的测试数据。比如 ImageNet 里面有 120 万张图片，是从 Flickr 等照片网站下载过来的。反反复复都是那些，所以实际的准确率和识别效果值得怀疑。数据全都是网络上的照片，但网络上数据肯定是不全面的，拍照的角度和光线都无法概括现实的多样性。而且不管是训练还是测试的数据，他们选择的都是在理想环境下的照片，没有考虑各种自然现象：反光，折射，阴影等。

比如下图就是图像识别常用的 ImageNet 和其它几个数据集的一小部分。你可以看到它们几乎全都是光线充足情况下拍的照片，训练和测试用的都是这样的照片，所以遇到现实的场景，光线不充足或者有阴影，准确率很可能就没有 paper 上那么高了。

如此衡量“准确率”，有点像你做个编译器，却只针对很小一个 benchmark 进行优化跑分。一旦遇到实际的代码，别人可能就发现性能不行。但神经网络训练需要的硬件等条件比较昂贵，一般人可能

也很少有机会进行完整的模型训练和实际的测试，所以大家只有任凭业内人士说“超人类准确率”，却无法验证它的实际效果。

“Top-5 准确率”的骗局

不但测试数据的“通用性”值得怀疑，所谓“准确率”的计算标准也来的蹊跷。AI 领域向公众宣扬神经网络准确率的时候，总喜欢暗地里使用所谓“top-5 准确率”，也就是说每张图片给 5 次机会分类，只要其中一个对了就算正确，然后计算准确率。依据 top-5 准确率，他们得出的结论是，某些神经网络模型识别图像的准确率已经“超越了人类”。

如果他们提到“top-5”还算好的了，大部分时候他们只说“准确率”，而不提“top-5”几个字。在跟人比较的时候，总是说“超越了人类”，而绝口不提“top-5”，不解释是按照什么标准。我为什么对 top-5 有如此强烈的异议呢？现在我来解释一下。

具体一点，“top-5”是什么意思呢？也就是说对于一张图片，你可以给出 5 个可能的分类，只要其中一个对了就算分类正确。比如图片上本来是汽车，我看到图片，说：

1. “那是苹果？”
2. “哦不对，是杯子？”
3. “还是不对，那是马？”
4. “还是不对，所以是手机？”
5. “居然还是不对，那我最后猜它是汽车！”

五次机会，我说出 5 个风马不及的词，其中一个对了，所以算我分类正确。荒谬吧？这样继续，给很多图片分类，然后统计你的“正确率”。

为什么要给 5 次机会呢？ImageNet 比赛（ILSVRC）对两种不同的比赛给出了两种不大一样的说法。一种说是为了让机器可以识别出图片上的多个物体，而不因为其中某个识别出的物体不是正确标签（ground truth）而被算作错误。另外一种说是为了避免输出意义相同的近义词，却不能完全匹配标签而被算作错误。

两个说法的理由不同，但数学定义基本是一样的。总之就是有五次机会，只要对了一个就算你对。

看似合理？然而这却是模糊而错误的标准。这使得神经网络可以给出像上面那样风马不及的 5 个标签（苹果，杯子，马，手机，汽车），其中前四个都不是图片上的物体，却仍然被判为正确。

你可能觉得我的例子太夸张了，但是准确率计算标准不应该含有这样的漏洞。只要标准有漏洞，肯定会有错误的情况会被放过。现在我们来看一个实际点的例子。

上图是一个 Coursera 的机器学习课程给出的 top-5 实际输出结果的例子。你可以从中发现，纵然有一些 top-5 输出标签是近义词，可是也有很多并不是近义词，而是根本错误的标签。比如“算盘”图片的 top-5 里面包含了computer keyboard（电脑键盘）和 accordion（手风琴）。“老虎”图片的 top-5 里面包含了两种狗的品种名字（boxer，Saint Bernard）。

另外你还可以看到，测试图片是经过精心挑选和裁剪的，里面很少有多于一个物体。所以第一种说法，“可能输出某个图片上存在的物体但却不是正确答案”，恐怕是很少见的。

所以 ILSVRC 对使用 top-5 给出的两个理由是站不住脚的。它想要

解决的问题并不是那么突出地存在，但是它却开了一道后门，可能放过很多的错误情况。比如上面的“算盘”图片，如果排名第一的不是 abacus，而是 computer keyboard（电脑键盘）或者 accordion（手风琴），只要 abacus 出现在 top-5 列表里，这个图也算识别正确。所以 top-5 根本就是错误的标准。

其实要解决图片上有多个物体的问题，或者输出是近义词的问题，都有更好的办法，而不会让错误的结果被算成正确的。每一个学过基础数据结构和算法的本科生都应该能想出更好的解决方案。比如你可以用一个近义词词典，只要输出的标签和“正确标签”是近义词就算正确。对于有多个物体的图片，你可以在标注时给它多个标签，算法给出的标签如果在这个“正确标签集合”里面就算正确。

但 ILSVRC 并没有采用这些解决方案，而是采用了 top-5。这么基础而重要的问题，AI 业界的解决方案如此幼稚，却被全世界研究者广泛接受。你们不觉得蹊跷吗？我觉得他们有自己的目的：top-5 使得神经网络的准确率显得很高，只有使用这个标准，神经网络才会看起来“超越了人类”。

Top-5 准确率总是比 top-1 高很多。高多少呢？比如 ResNet-152 的 top-1 错误率是 19.38%，而 top-5 错误率却只有 4.49%。Top-1 准确率只能算“勉强能用”，换成 top-5 之后，忽然就可以宣称“超越人类”了，因为据说人类的 top-5 错误率大概是 5.1%。

Top-5 准确率对人是不公平的

可能很多人还没意识到，top-5 比较方法对人是不公平的。图片上要是人见过的物体，几乎总是一次就能做对，根本不需要 5 次机会。使用“top-5 准确率”，就像考试的时候给差等生和优等生各自 5 次机会来做对题目。当然，这样你就分不清谁是差等生，谁是优等生了。“top-5 准确率”大大的模糊了好与坏之间的界线，最后看起来都差不多了，甚至差等生显得比优等生还要好。

具体一点。假设一个人识别那些图片的时候，他的 top-5 错误率是 5.1%（就像他们给出的数字那样），那么他的 top-1 错误率大概也是 5.1%。因为人要是一次机会做不对，那他可能根本就没见过图片上的物体。如果他一次做不对，你给他 5 次机会，他也做不对，因为他根本就不知道那东西叫什么名字。

现在某个神经网络（ResNet-152）的 top-5 错误率是 4.49%，它的 top-1 错误率是 19.38%。你却只根据 top-5 得出结论，说神经网络超越了人类。是不是很荒谬？

退一万步讲，就算你可以用 top-5，像这种 4.49% 与 5.1% 的差别，只相差 0.61%，也应该是忽略不计的。因为实验都是有误差，有随机性的，根据测试数据的不同也有差异，像这样的实验，1% 以内的差别根本不能说明问题。如果你仔细观察各个文献列出来识别率，就会发现它们列出的数字都不大一样。同样的模型，准确率差距可以有 3% 以上。但他们拿神经网络跟人比，却总是拿神经网络最好的那个数，跟人死扣那百分之零点几的“优势”，然后欢天喜地宣称已经“超人类”了。

而且他们真的拿人做过公平的实验吗？为什么从来没有发布过“神经网络 vs 人类 top-1 对比结果”呢？5.1% 的“人类 top-5 准确率”数字是哪里来的呢？哪些人参加了这个测试，他们都是什么人？我唯一看到对人类表现的描述，是在 Andrej Karpathy 的主页上。他拿 ImageNet 测试了自己的识别准确率，发现好多东西根本没见过，不认识，所以他又看 ImageNet 的图片“训练”自己，再次进行测试，结果准确率大大提高。

就那么一个人得出的“准确率”，就能代表全人类吗？而且你们知道 Andrej Karpathy 是谁吧。他是李飞飞的学生，目前是 Tesla 的 AI 主管，而李飞飞是 ImageNet 的发起者和创造者。让一个“内幕人士”拿自己来测试，这不像是公正和科学的实验方法。你见过有医学家，心理学家拿自己做个实验，就发表结果的吗？第一，人数太少，至少应该有几十个智商正常的人来做这个，然后数据平均一下吧？第二，这个人是个内幕人士，他的表现恐怕不具有客观性。

别误会了，我并不否认 Andrej Karpathy 是个很聪明，说话挺耿直的人。我很欣赏他讲的斯坦福 cs231n 课程，通过他的讲述我第一次明白了神经网络到底是什么，明白了 back-propagation 到底如何工作。我也感谢李飞飞准备了这门课，并且把它无私地放在网上。但是这么大一个领域，这么多人，要提出“超越了人类视觉”这么大一个口号，居然只有研究者自己一个人挺身而出做了实验，你不觉得这有点不负责任吗？

AI 领域对神经网络训练进行各种优化，甚至专门针对 top-5 进行优化，把机器的每一点性能每一点精度都想榨干了去，对于如何让人准确显示自己的识别能力，却漫不经心，没有组织过可靠的实验，准确率数字都不知道是怎么来的。对比一下生物，神经科学，医学，这些领域是如何拿人做实验，如何向大家汇报结果，AI 领域的做法像是科学的吗？

这就是“AI 图像识别超越人类”这种说法来的来源。AI 业界所谓“超人类的识别率”，“90+% 的准确率”，全都是用“top-5 准确率”为标准的，而且用来比较的人类识别率的数字没有可靠的来源。等你

用“top-1 准确率”或者更加公平的标准，使用客观公正抽选的人类实验者的时候，恐怕就会发现机器的准确率远远不如人类。

尴尬的 top-1 准确率

我们来看看 top-1 准确率吧。业界最先进的模型之一 ResNet-152 的 top-1 错误率是 19.38%。2017 年的 ImageNet 分类冠军 SENet-154，top-1 错误率是 18.68%。当然这也没有考虑过任何实际的光线，阴影和扭曲问题，只是拿标准的，理想情况的 ImageNet “测试图片”来进行。遇到实际的情况，准确率肯定会更低。

神经网络要想提高 top-1 准确率已经非常困难了，都在 80% 左右徘徊。有些算法工程师告诉我，识别率好像已经到了瓶颈，扩大模型的规模才能提高一点点。可是更大的模型具有更多的参数，也就需要更大规模的计算能力来训练。比如 SENet-154 尺寸是 ResNet-152 的 1.7 倍，ResNet-152 尺寸又是 ResNet-50 的 2.4 倍，top-1 准确率才提高一点点。

我还有一个有趣的发现。如果你算一下 ResNet-50 和 ResNet-152 的差距，就会发现 ResNet-152 虽然模型大小是 ResNet-50 的 2.4 倍，它的 top-1 错误率绝对值却只降低了 1.03%。从 22.37% 降低到 21.34%，相对降低了 (22.37-21.24)/22.37 = 4.6%，很少。可是如果你看它的 top-5 错误率，就会觉得它好了不少，因为它从 6.36% 降低到了 5.54%，虽然绝对值只少了 0.82%，比 top-1 错误率的改进还小，可是相对值却降低了 (6.36-5.54)/6.36 = 12.9%，就显得改进了挺多。

这也许就是为什么 AI 业界用 top-5 的第二个原因。因为它的错误率基数很小，所以你减小一点点，相对的“改进”就显得很多了。然后你看历年对 top-5 的改进，就觉得神经网络识别率取得了长足的进步！

而如果你看 top-1 准确率，就会觉得几乎没有变化。模型虽然大了几倍，计算量大了那么多，top-1 准确率却几乎没有变。所以神经网络的 top-1 准确率似乎确实到了一个瓶颈，如果没有本质的突破，恐怕再大的模型也难以超越人类。

AI 业界的诚信问题和自动驾驶的闹剧

准确率不够高，不如人类其实问题不大，只要你承认它的局限性，把它用到能用的地方就行了。可是最严重的问题是人的诚信，AI 人士总是夸大图像识别的效果，把它推向超出自己能力的应用。

AI 业界从来没有向公众说清楚他们所谓的“超人类识别率”是基于什么标准，反而在各种媒体宣称“AI 已经超越了人类视觉”。这完全是在欺骗和误导公众。上面 Geoffrey Hinton 的采访视频中，主持人也提到“神经网络视觉超越了人类”，这位深度学习的先驱者对此没有任何说明，而是欣然接受，继续自豪地夸夸其谈。

你可以给自动驾驶车 5 次机会来判断前面出现的是什么物体吗？你有几条命可以给它试验呢？Tesla 的 Autopilot 系统可能 top-5 正确率很高吧：“那是个白板…… 哦不对，那是辆卡车！”“那是块面包…… 哦不对，那是高速公路的隔离带！”

我不是开玩笑，你点击上面的“卡车”和“隔离带”两个链接，它们指向的是 Tesla Autopilot 引起的两次致命车祸。第一次车祸，Autopilot 把卡车识别为白板，直接从侧面撞上去，导致车主立即死亡。另一次，它开出车道，没能识别出高速公路中间的隔离带，完全没有减速，反而加速撞上去，导致车主死亡，并且着火爆炸。

神经网络能把卡车识别为白板还算“top-5 分类正确”，Autopilot 根本没有视觉理解能力，这就是为什么会引起这样可怕的事故。

你可以在这里看到一个 Autopilot 导致的事故列表。

出了挺多人命，可是“自动驾驶”的研究仍然在混沌中进行。2018 年 3 月，Uber 的自动驾驶车在亚利桑那州撞死一名推自行车过马路的女性。事故发生时的车载录像已经被公布到了网上。

报告显示，Uber 的自动驾驶系统在出事前 6 秒钟检测到了这位女士，起初把她分类为“不明物体”，然后分类为“汽车”，最后分类为“自行车”，完全没有刹车，以每小时 40 英里的速度直接撞了上去…… 【新闻链接】

在此之前，Uber 被加州政府吊销了自动驾驶实验执照，后来他们转向了亚利桑那州，因为亚利桑那州长热情地给放宽政策，“拥抱高科技创新”。结果呢，搞出人命来了。美国人看到 Uber 自动车撞死人，都在评论说，要实验自动驾驶车就去亚利桑那州吧，因为那里的人命不值钱，撞死不用负责！

据 2018 年 12 月消息，Uber 想要重新开始自动驾驶实验，这次是在宾夕法尼亚州的匹兹堡。他们想要在匹兹堡的闹市区进行自动驾驶实验，因为那里有狭窄的街道，列车铁轨，许多的行人…… 我觉得要是他们真去那里实验，可能有更好的戏看了。

自动驾驶领域使用的视觉技术是根本不可靠的，给其它驾驶者和行人造成生命威胁，各个自动驾驶公司却吵着想让政府交通部门给他们大开绿灯。某些公司被美国政府拒绝批准牌照之后大吵大闹，骂政府监管部门不懂他们的“高科技”，太保守，跟不上时代。有的公司更是异想天开，想要政府批准他们的自动车上不安装方向盘，油门和刹车，号称自己的车已经不需要人类驾驶员，甚至说“只有完全去掉了人类的控制，自动车才能安全运行。”

一出出的闹剧上演，演得好像自动驾驶就快实现了，大家都在拼命抢夺这个市场似的，催促政府放宽政策。很是有些我们当年大炼钢铁，

超英赶美的架势。这些公司就跟小孩子耍脾气要买玩具一样，全都吵着要爸妈让他玩自动驾驶，各种蛮横要求，马上给我，不然你就是不懂高科技，你就是“反智”，“反 AI”，你就是阻碍历史进步！给监管机构扣各种帽子，却完全不理解里面的难度，伦理和责任。玩死了人，却又抬出各种借口，不想负责任。

虽然 Tesla 和 Uber 是应该被谴责的，但这里面的视觉问题不只是这两家公司的问题，整个自动驾驶的领域都建立在虚浮的基础上。我们应该清楚地认识到，现有的所谓 AI 根本没有像人类一样的视觉理解能力，它们只是非常粗糙的图像识别，识别率还远远达不到人类的水平，所以根本就不可能实现自动驾驶。

什么 L1~L4 的自动驾驶分级，都是瞎扯。根本没法实现的东西，分了级又有什么用呢？只是拿给这些公司用来忽悠大家的口号，外加推脱责任的借口而已。出事故前拿来做宣传：“我们已经实现 L2 自动驾驶，目前在研究 L3 自动驾驶，成功之后我们向 L4 进军！” 出事故后拿来推脱责任：“我们只是 L2 自动驾驶，所以这次事故是理所当然，不可避免的！”

如果没有视觉理解，依赖于图像识别技术的“自动驾驶车”，是不可能在复杂的情况下做出正确操作，保障人们安全的。机器人等一系列技术，也只能停留在固定场景，精确定位的“工业机器人”阶段，而不能在复杂的自然环境中行动。

识别技术还是有意义的

要实现真正的语言理解和视觉理解是非常困难的，可以说是毫无头绪。一代又一代的神经学家，认知科学家，哲学家，为了弄明白人类“认知”和“理解”到底是怎么回事，已经付出了许多的努力。可是直到现在，对于人类认知和理解的认识都不足以让机器具有真正的理解能力。

真正的 AI 其实没有起步，很多跟 AI 沾点边的人都忙着忽悠和布道，没人关心其中的本质，又何谈实现呢？除非真正有人关心到问题所在，去研究本质的问题，否则实现真的理解能力就只是空中楼阁。我只是提醒大家不要盲目乐观，不要被忽悠了。与其夸大其词，欺骗大众，说人工智能快要实现了，不如拿已有的识别技术来做一些有用的事情，诚实地面对这些严重的局限性。

我并不是一味否定识别技术，我只是反对把“识别”夸大为“理解”，把它等同于“智能”，进行不实宣传，用于超出它能力的领域。诚实地使用识别技术还是有用的，而且蛮有趣。我们可以用这些东西来做一些很有用的工具，辅助我们进行一些事情。从语音识别，语音合成，图片搜索，内容推荐，商业金融数据分析，反洗钱，公安侦查，医学图像分析，疾病预测，网络攻击监测，各种娱乐性质的 app…… 它确实可以给我们带来挺多好处，实现我们以前做不到的一些事情。

另外虽然各公司都在对他们的“AI 对话系统”进行夸大和不实宣传，可是如果我们放弃“真正的对话”，坦诚地承认它们并不是真正的在对话，并没有智能，那它们确实可以给人带来一些便利。现有的所谓对话系统，比如 Siri，Alexa，基本可以被看作是语音控制的命令行工具。你说一句话，机器就挑出其中的关键字，执行一条命令。这虽然不是有意义的对话，却可以提供一些方便。特别是在开车不方便看屏幕的时候，语音控制“下一首歌”，“空调风量小一点”，“导航到最近的加油站”之类的命令，还是有用的。

但不要忘记，识别技术不是真的智能，它没有理解能力，不能用在自动驾驶，自动客服，送外卖，保洁阿姨，厨师，发型师，运动员等需要真正“视觉理解”或者“语言理解”能力的领域，更不能期望它们取代教师，程序员，科学家等需要高级知识的工作。机器也没有感情和创造力，不能取代艺术家，作家，电影导演。所有跟你说机器也能有“感情”或者“创造力”的都是忽悠，就像现在的对话系统一样，只是让人以为它们有那些功能，而其实根本就没有。

你也许会发现，机器学习很适合用来做那些不直观，人看不透，或者看起来很累的领域，比如各种数据分析。实际上那些就是统计学一直以来想解决的问题。可是视觉这种人类和高等动物的日常功能，机器的确非常难以超越。如果机器学习领域放弃对“人类级别智能”的盲目追求，停止拿“超人类视觉”一类的幌子来愚弄大众，各种夸大，那么他们应该能在很多方向做出积极的贡献。

（全文完）