国际 科技 深度

人脸识别到底是甚么？演算法是不是无所不能？

人脸识别系统无处不在，我们似乎已无法抗拒一举一动被记录下来；但这些模拟人脑的演算法，其实远非牢不可破。

2018年10月，中国北京一个保安科技博览展。摄：Thomas Peter/Reuters/达志影像

朱孝文

特約撰稿人 朱孝文 發自倫敦 | 2022-09-06

人面识别　　人脸识别　　科技

〔编者按〕2019年，BBC的一套记录片拍到了伦敦警察试用人脸识别系统的情景：一名（貌似有中东血统的）男子在摄像头前把衣领拉高，结果被警察截停审问。行人质疑警察截查该男子的理由，警察说是因为“他看到镜头就立即尝试遮脸”，所以他们认为他形迹可疑。这个不想让摄像头拍到自己的男子，最终还是被警察拍了照。

人脸识别不止在中俄等威权国家蓬勃发展，即使在民主社会，这种技术还是引起了许多关于私隐、人权的争议，而且社会对个体的保障似乎远没有科技发展那么迅速。端传媒九月将刊出一系列专题报道，探讨中国和俄罗斯成为监控社会之路。这篇文章是系列的第一篇，将解释人脸识别的原理，以及个体“逃离”识别系统在技术上的可能性。请持续留意端传媒报道。

（朱孝文，人工智能研究員）

她打开了自己的相簿，把一张自己早两天拍下的生活照，从桌面上拉到网站的搜寻框。不消数十秒，网站显示了一堆搜寻结果：那是从互联网上九亿张图片中找出来的她的照片，里面有一些连她自己也未看过，有些她甚至已经戴了口罩或太阳眼镜。更甚的是，里面一些图片来自色情网站，虽然主角并不是她本人。这是《纽约时报》编辑Kashmir Hill在2022年5月测试一个名为PimEyes的人脸识别搜寻网站时的经历。我让端的编辑看看她的照片在PimEyes显示了甚么结果，她除了看到自己的照片，搜寻结果中也有跟她相似的陌生人（而且也有非她本人并来自色情网站的照片，正如Hill所言，单是系统认为这“可能是她们”，就已经够令人不安了）。

十多年前谷歌Google Photo推出试用服务，大概是我第一次见识人脸识别系统潜力的契机——我将数千张照片拉到平台上，系统自动将数百个被拍摄者辨认出来。最让我惊讶的，是同一个人跨越不同年龄段的照片都被它认出来了，例如童年时期的我与大学时期的我，Google的系统都能认作同一个人（虽然有少量我的照片被它辨识作另外两个“我”，特别是侧面照），在当时确实非常厉害。现在人面辨识已在各种应用上无处不在，从网上申请银行帐户时所提交的自拍所触发的“电子模式认识你的客户”（Electronic Know Your Customer，eKYC）验证，到辨理签证及各国机场入闸时的资料核对，都用上了这种技术。

过去十年，随著深度学习的发展，人脸辨识技术也愈来愈成熟。现在它能从茫茫人海中将个人准确锁定，如果被没有权力制衡的组织使用，将会是我们每个人的恶梦。人脸识别系统自2010年代晚期开始普及进北京社区，根据外媒报道，这个系统也在不同城市扫描居民面孔，以判断是否有维族人。配合社会信用体系，人脸辨识能令个人私生活无所遁形。2022年初，国际特赦组织也发布了研究报告，指纽约警方在黑人和拉丁裔社区的许多天眼有人脸辨识功能，是对个人自由的侵犯。而近年大行其道却名声败坏的DeepFake演算法，也与个人私隐息息相关。

人面辨识系统如此强大，是不是就完全没法抗衡？起码现在还不是——在许多研究中，输入一些经过加工的图像会造成辨识系统误判，将图中人物辨认作另一个随机或指定的人。这些加工图片也可以让看到人脸

的图像会造成辨识系统误判，将图中人物辨认作为一个随机或指定的人。这些加上图片也可以让看到人脸图片的系统无法辨认出有人脸的存在。这些研究，对于个人面对人面辨识系统时的可能性，或可带来一些启示。

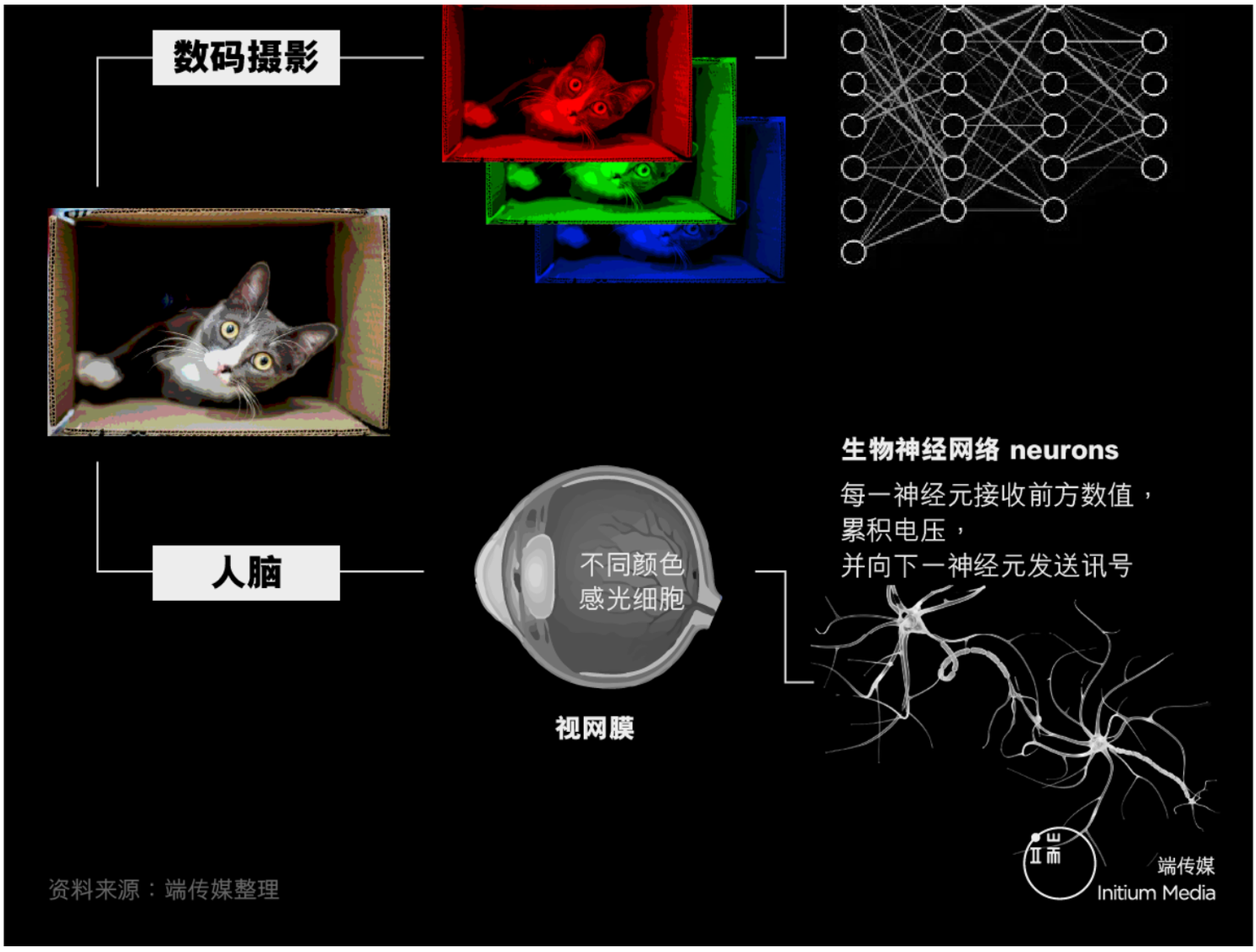
甚么是“人工智能”？

要明白为何人面识别系统并非坚不可摧，我们需要了解它如何运作。对辨识系统进行除魅，也可让我们不会过份渲染它的强大，能以更理性的角度去审视它的能与不能。

首先，人脸是一种数据：有时以图片方式表达（计算机处理二维照片的传统方式是三原色的二维矩阵），有时以“重要地标”坐标位置表达（选择性的点云（Point cloud），演算法处理特定三维模型的一种方法）的数据。无论是用甚么方式表达，要处理数据就要用到统计学。

人脸辨识系统是机械学习（Machine Learning）演算法的一种。人工智能领域从80年代规则导向（rule based）的专家系统开始，发展到今日数据导向懂得自我调节的机械学习（Machine Learning）演算法，后者的理论基础就是源于统计学。机械学习演算法透过用家给予的数据以调整自己计算时所使用的参数（也就是机械学习的“学习”），从而对统计数据的概率分布（distribution）进行模拟。因此人脸辨识系统需要数据学习，越多的数据，越能让更复杂的机械学习模型有效学习。在这情况下，能以较低成本收集并使用大量人面数据的国家或企业，就越有能力开发更强大的人脸辨识系统。现时在全球人脸识别演算法方面排名前列的，几乎都是美国、中国和俄罗斯的企业。

今日主流的人脸辨识系统，属于机械学习中近10年来进展迅速的深度学习（Deep Learning）子分类。这是将计算机的计算单元以一种层层递进方式连结起来的演算法，这种计算结构参考了人脑神经元的运作方式（这也是为何计算单元也被称作“人工神经元”（artificial neuron），而深度学习模型有时也被称作“人工神经网络”）。人脑是宇宙间最复杂的组织，虽然我们还是认为它是单纯的物质，但它实际如何运作，无论是脑神经科学家还是人工智能科学家，到现在也没能完全掌握得到。人工智能之父，著名电脑科学家候世达（Douglas Hofstadter）认为人脑是“不可化约（irreducible）”的——即是我们无法有效地用更基本的语言归纳它。但现在人工神经网络的运作方式，已经是现有理解中比较能模拟人脑运作的方式了。



也因为它需要极复杂的神经元连接来模拟人脑，这种演算法需要的计算资源非常庞大。所以虽然它早在80年代末诞生，但在2010年代前主要只出现在学论研究中。过去十多年，由于计量资源以及收集数据成本的下降，深度学习的研究与应用迅速成长。其中一个里程碑，是深度学习模型AlexNet在2012年的ImageNet挑战赛（参加者需要在将巨量图片按指定的1000类别分类，例如犬只就有90个类别）中，击败一众依赖传统统计学与计算机视觉演算法的对手，成为史上第一个赢得此比赛的深度学习模型。

AlexNet与很多其它后来的图片分类模型，都属于一种称为卷积神经网络（Convolutional Neural Network, CNN）的结构。“卷积”是电子工程中一种传统进行微积分的方法，原理可概括为拿取一个特定区域中数据加权平均的值。CNN中二维卷积配合深度学习的灵感，源自灵长类动物视觉系统从视网膜连到视皮质（Visual cortex）连接路径的运作方式。一开始的“卷积”计算能让网络了解图片的区域资讯，例如图片的某一部份是否开始出现物件边界，物件的颜色／纹理等等；而透过深度学习连结，在深层的计算单元能得出更抽象的知识（例如物件是否有“眼睛”，是否“动物”等等），最后作出分类。

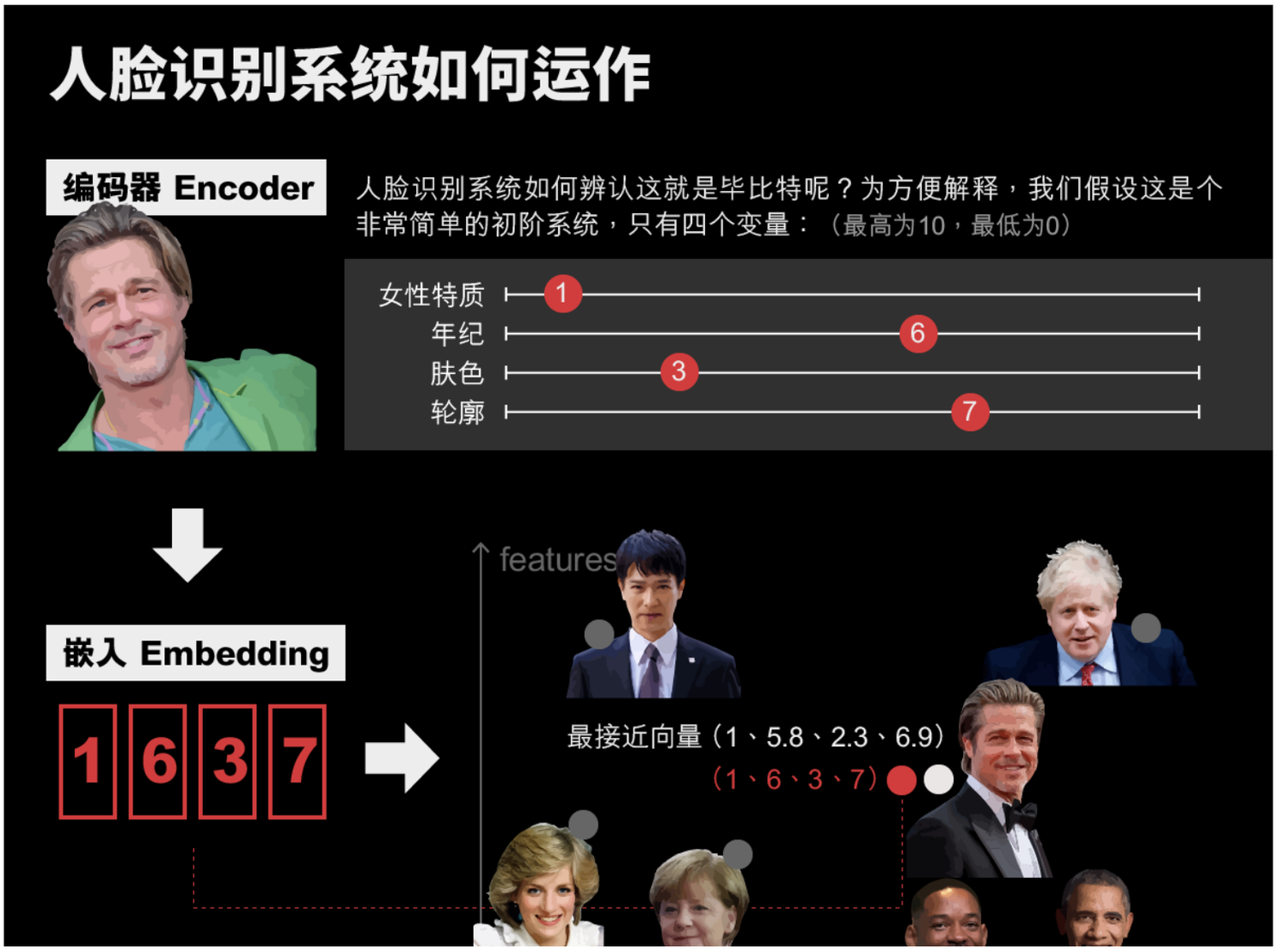
人脸识别是怎么把你认出来的？

而典型的人工智能是如何将图像分类的呢？如果预设了有1000个分类（如ImageNet挑战赛），那么这个系统的输出，就是一个有1000个数字的向量（vector）；1000个数字加起来刚好等于1，分别代表1000个类型各自的概率。假设第539个数字属于熊猫分类，而这数字的数值是80%，这代表了演算法认为该图片有80%的概率为熊猫。

问题来了，地球有七十亿人口，每个人都长得不一样，难道要有七十亿个分类？这种做法当然不太理想：第一，用分类来输出太巨量，也很没效率。第二，与分类系统不同，很多我们要让人面辨识系统辨认的面孔，可以是事前它从未见过的。例如前述Google Photo的例子：系统以前还没有看过你的照片，但当你把一堆自己的相片拉进平台，它就懂得自己分辨不同图片里的人是否同一人。

所以人脸需要一种新的表达方式。描述得太仔细不一定是好事：点云能记录面部每一个重要地标的三维坐标，但同一个人只要表情不同，点云也会不一样（当你笑的时候，嘴角的坐标就被拉到另一处了）。我们希望这种面孔的表达，可以代表一个人，而不是代表这个人的特定表情，同时也能表达系统未见过的人。

所以，我们需要一种叫“嵌入”（embedding）的表达方法。



如果有个监视器拍到影星毕彼特（Brad Pitt；另译：毕比特、布莱德·彼特），系统怎么知道他就是毕彼特呢？为方便解释，我们假设这个是非常简单的初阶系统，只有四个变量：女性特质、年纪、肤色和轮廓。长得愈阴柔、圆润的人，“女性特质”的变量也就愈高（最高为10，最低为0），轮廓愈深，“轮廓”的变量也愈高，如此类推。以毕彼特为例，他的女性特质可能是1，中年所以年纪是6，肤色较白所以肤色是3，轮廓深所以最后一项是7——那他的“嵌入”向量就是（1, 6, 3, 7）。假设系统也拍到了米雪儿．奥巴马（Michelle Obama），她的变量则可能是（8, 6, 6, 6），拍到我的话，我的变量则可能是（3, 3, 5, 4）。

演算法拿到嵌入向量后，就会拿到数据库与预先储存起来的嵌入向量比较，发觉与眼前人的嵌入（1, 6, 3, 7）最接近的向量是（1, 5.8, 2.3, 6.9），而（1, 5.8, 2.3, 6.9）属于毕彼特，因此系统就会判断目前人是毕彼特了。当然我们早就脱离了规则导向的年代，不用自己决定这些变量，只要有大量数据（即人脸）让系统学习，它会自动决定那些变量是甚么。

嵌入是很多人面辨识系统（如前面提到的谷歌FaceNet模型，以及脸书2014年的DeepFace模型等等）的典型输出。我们试图用少量的数值（低维度），去表达原本甚为庞大的物件（高维度），所以嵌入是一个降维（dimensionality reduction）的过程。例如人脸重要地标的坐标（上文提到的点云）总共可以有几万个数字，但其实用较少的数字去代表脸孔是可能的。例如2015年谷歌FaceNet模型的嵌入，就用上4,975个数字（也就是4,975维）去代表一个面孔。我曾在人脸嵌入开展过一些研究工作，当时使用的理论根基是2016年Anh Tuan Tran等学者发展出来的99维嵌入，然后将每张图片的99维嵌入还原为原始的人脸。还原的方法一般是使用一些生成演算法，例如近年颇热门的GAN（Generative Adversarial Network）。

这么少的嵌入数值，之所以足够“大”去承载原本更庞大的数据，是因为机械学习中的流形假设（manifold hypothesis）信念：很多表面上高维度的物体数据结构其实十分“松散”，他们可以用一种紧密得多的低维度表达方式（低维流形）去表达。嵌入一词源自数学中的拓扑学（Topology），用来判断两个空间是否等价。人工智能领域借用此词时，则用来判断两个数据空间是否“足够近似”，假如人脸嵌入大约能将原始人脸压缩，并在将来还原，那么嵌入在某程度上与原始人脸就是“足够近似”了。

常见的嵌入除了人面辨识的面孔嵌入（facial embedding）外，自然语言处理（Natural language processing，NLP）的演算法输出的文字嵌入（word embedding）也是另一例子。一个有趣的例子是“国王”与“女皇”两词两个文字嵌入之间的距离，与“男人”与“女人”两词两个文字嵌入之间的距离颇为接近。说到这里，我们大概都可以看到，嵌入在某程度上就好像人类对事物的认知（perception），也就是演算法对知识的一种理解。深度学习中一种经典的自编码器（autoencoder）模型，就是将原始图片转化作低维的嵌入，再由嵌入重新转成图片，然后研究人员会比较输出与输入的差异有多少，以判断嵌入是否能有效“储存”原始资讯，达到流形假设所期待的效果。

所以“嵌入”与“分类”两种输出本质十分不同，前者是特定物件重要特征的表达，后者纯粹是将物品划入已预先设定好的分类。如前所述，嵌入所代表的意义并不一定是明确的，这些数值只是演算法透过数据所学习到的表达方式。例如不一定像之前毕彼特的例子一样，有一个指定的数值去决定女性特质、年纪、肤色和轮廓等等。

英国伦敦街头，警方展示一张告示牌指该区正在测试人脸识别技术。摄：Kirsty O'Connor/PA Images via Getty Images

人面辨识模型会将不同照片的嵌入数值分群，数值相近的嵌入会被列入一个相同的群（cluster）。也就是

说，同一个人的不同照片，我们希望其嵌入数值彼此相近。因此要愚弄人面辨识模型的话，意思其实是要操弄演算法的嵌入空间，希望被“添加”了的输入，能让演算法将图片的嵌入投放至远处，令它无法成功识别拍到的人。如果我们想愚弄的不单单是一个演算法，而是同一时间愚弄多个不同的人面辨识系统，操弄嵌入空间就变得更加重要。根据上文提及的流形假设，人面资讯在不同演算法下的“本质维度”应该要是近似的，不同演算法的嵌入亦很有可能有相似的几何结构，因此只要能操弄一个演算法的嵌入，照理这种操弄亦应能转移到其它演算法上，并成功欺瞒它们。

骗不了人，但却能骗过强大的演算法

两年前，网络安全公司McAfee的团队尝试攻击机场用于验证护照的面部识别系统。他们用机器学习生成了在人眼里看起来像人，但机器却“看不懂”的图像，并成功欺骗了人面识别系统。研究人员用的是一个叫CycleGAN的图像翻译算法，这种演算法擅长将照片改变风格，例如将马的影片变成班马影片，或是将夏季绿油油的山脉照片，变成冬季的积雪山脉。McAfee团队将两个人（甲和乙）的图像输入CycleGAN，让它帮两张照片互相“转换风格”，最终生成了一张人类看起来像甲，但机器却会以为是乙的图像。所以如果甲在禁飞名单上，这种技术会让他避过人面识别系统，成功上机。

有GAN之父之称的Goodfellow，在2014年就指出卷积神经网络其实很容易被欺瞒。一张熊猫图片加上了一些精心计算的扰动（perturbation），在人眼看起来图片几乎毫无改变，依旧是一张熊猫图片，但当时最先进的卷积模型会将图片以99%信心度误判为黑猩猩。这种用作攻击的图片，在研究领域被称作对抗性例子（Adversarial Example）。但这种程度的扰动，对人类大脑而言是完全没用的，没有人会怀疑这不是熊猫。事实上，要人眼不察觉图片被扰动修改，是对抗性例子中一个很常见的要求。这就是所谓的“隐闭性”－－让图片不会被系统的人类维护员发现“有诈”。这也说明了，参考人类皮质运作方式而创造的卷积模型结构，与灵长类动物的视觉系统始终有巨大差别。

在这个方向发展，一些攻击演算法开始能制造出能欺瞒人面辨识系统的图片，例如Zhang等人在2020年论文中对中国腾讯的人面辨识系统展开攻击，攻击图片的面孔在人类眼中就与原图片无异，但辨识系统会错误地将图中人物身份误判为另一人。这些误导的目标可以是随机的，也可以是一个特定的目标。特定的目标当然较随机目标困难，但并非不可能，例如私人企业ADVERSA他们为客户测试系统健壮性（robustness）时所提供的攻击，就可以将照片面孔误导往指定人物（他们在广告中，就示范如何轻易让系统以为眼前人是SpaceX创始人马斯克（Elon Musk））。

这是一个很有趣的效果：系统的“眼睛”明明看到一个分明不是马斯克的人，但它大脑看到的却是马斯克。这又跟吃了迷幻药，然后见到不存在事物的人类不一样－－我们可以控制系统见到甚么，却没有一颗迷幻药能控制嗑药的人见到耶稣还是佛祖。除了将图片辨识成另一身份，对抗性图像也可以使系统看著前方人面的时候辨识不到人脸的存在，例如2016年Sharif等人的研究便让人物带上花纹眼镜以达到“隐身”之效，跟漫画《多啦A梦》里一只要戴著就能让人视而不见的石头帽差不多。

跟漫画《多啦A梦》里，只要戴着就能让人视而不见的路人帽差不多。

英国伦敦街头一部闭路电视。摄：Dominika Zarzycka/NurPhoto via Getty Images

当然，当面对真实的摄影镜头时，要扰乱系统就比静态图像要困难得多，但也还不是没可能。早在2001年，Viola-Jones演算法（较低层次，准确率较低的非CNN人脸识别演算法）出现时，就有像Computer Vision Dazzle这样的反人脸识别化妆出现。当然如果不是身在日本秋叶原，化个这样的妆外出会非常突兀，而事实上这些意图扰乱光暗的化妆，已经骗不过今日的CNN算法。

但对抗性例子仍然不断出现：在2018年Brown等人的研究中，他们仅仅是在香蕉旁边放下了一块贴有扰动图片的杯垫（他们称之为adversarial patch；对抗性补丁），就可让实时的摄影镜头将香蕉错误分類为多士炉。谷歌在2017年制作了一只海龟玩偶，在实时的摄影镜头监察下总会被分类为步枪。这些都是在物理环境下的对抗性例子。

当然 – – 防卫者（也就是辨识系统的维护人员）也有辨法将高墙加高，例如使用更多的数据对系统进行训练，或者使用训练数据前先将图片做平移、扭曲或加噪等的数据扩充（Data Augmentation）步骤，甚至在进行辨识前先额外用传统方式对照片进行消噪等等。但无论如何，对系统的成功攻击让我们看到，强大的辨识系统并非毫无破绽。

虽然我们谁都不知道人脸识别演算法之后的发展会如何，但观乎目前情况，攻击人脸识别系统的方法持续出现，亦没有人能设计出足够健壮的系统去抵抗各种攻击的方法。而目前很多提升了防御的系统，都有演算法判断力退化的问题。而对于物理环境的攻击，目前的防御都不是运作得很好。而在理论上，由于我们缺乏足够好的理论工具去描述“攻击——防御问题”的优化解，无人知道理论上是否能建立一个能防御各种攻击的系统。无疑人脸识别在许多国家的广泛使用令人担忧，但在技术上，人脸识别演算法离“牢不可破”还有一段距离。