经济观察报/2020 年/9 月/7 日/第 024 版
商业评论

应对“换脸”危机

商汤智能产业研究院战略生态研究主任　杨燕

继 2012 年“互联网+”后，“AI+”成为时代主旋律。然而在隐秘的角落，由 AI 所引发的安全风险和“黑灰产”问题正与日俱增。尤其是人脸识别——作为 AI 技术落地最广泛的场景之一，所面临的安全、伦理和道德等挑战愈发严峻。

经过近些年的快速发展，人脸识别已和智能硬件解锁、支付，以及公共服务等身份验证直接绑定在一起。因面部信息的唯一性，以及作为个人隐私中最敏感、重要的组成部分，一旦出现问题，将会对个人隐私、公共安全造成巨大威胁，故对技术的安全要求和标准相对更高。

以下，我们通过两个较为典型案例，来说一说 AI 攻防对人脸识别/人工智能技术和行业发展的作用和意义。

“换脸”技术的攻防战

2018 年，一段“奥巴马”呛声特朗普的视频在全美疯传。事后，这个视频被证明为伪造，其背后所利用的即是 AI“换脸”技术。该技术是基于生成对抗网络（GANs），通过两个模型——一个负责生成伪图，另一个负责鉴别伪图，对抗博弈的方式不断进化，从而达到以假乱真的水平。

而在更早的 2017 年，一位名为 deepfakes 的网友将色情电影中演员的脸替换成好莱坞女星，并将合成视频在 Reddit 网站上发布，引发全球热议，遭到大众对技术滥用的质疑。

这也为 AI“换脸”技术吸引了一波关注，DeepFake（深度伪造或深度合成）就此成为该技术的代名词，同名算法也在 Github 上开源，导致合成视频片段大量涌现。

根据创业公司 Deeptrace 报告显示，2019 年初，互联网上流转的、利用 DeepFake 技术生成的视频，共有 7964 个,仅仅 9 个月后，这个数字跃升至 14678 个，而其中就有高达 96%的 DeepFake 视频与色情相关。

DeepFake 技术的滥用引发全球担忧，也为人脸识别技术的应用推广带来了巨大风险。据称，2019 年底，硅谷人工智能公司 Kneron 曾使用 DeepFake 技术成功欺骗了支付宝和微信支付，并且顺利通过机场、火车站等自助终端检验。

虽然各国纷纷加强了监管措施，譬如美国政府公布了《禁止恶意深度伪造法案》《2019 年深度伪造责任法案》《2019 年深度伪造报告法案》，旨在通过限制 DeepFake 合成技术，打击虚假信息的传播。但遏制 DeepFake 技术滥用的根本手段，还是需要从安全对抗的本质上出发，铸造更高门槛的防御技术，以 AI 应对 AI，在攻防“互殴”之中不断增强系统的鲁棒性。

为此，科技巨头们也纷纷加入了这场“安全对抗”的战役之中。去年 9 月，谷歌开源了包含 3000 个 AI 生成的视频数据库，以支持社区加速开发 DeepFake 检测工具，对抗技术滥用风险；Facebook 也在同年 12 月发布了一套“反识别”系统，帮助辨别实时影像的真伪。

技术是中立的，也是双向发展的，不会因为惧怕风险而停滞不前。不久前，英伟达的研究人员提出了一种新的生成器架构，可基于风格迁移，将面部细节分离出来，并由模型进行单独调整，生成的面部图像比基于传统 GAN 技术更加逼真。

可见，假脸生成和真脸识别的算法对抗将会是持续的、动态的过程。

对抗样本的“攻防战”

第二个案例是，在 2019 年的世界黑帽安全大会上，腾讯的研究人员向与会者展示了如何利用一款缠着黑白胶带的眼镜，就能解锁苹果 FaceID。

如果这款技术还需要“受害人”的被动配合——趁“受害人”睡着，将眼镜戴在“受害人”

脸上，那么最新的 AI 技术，只需打印一张带有图案的纸条贴在脑门上，就能“戳瞎”AI 识别系统。

这种黑科技的应用，在学术上被称为“对抗样本”。百度百科对它的定义是“在数据集中通过故意添加细微的干扰所形成的输入样本，导致模型以高置信度给出一个错误的输出。”

目前大多数机器学习的鲁棒性都比较差，容易受对抗样本的影响。样本经过轻微修改后，输出结果可能会谬以千里，而且这些细微的修改人类几乎无法肉眼识别。

2018 年，在 GeekPwn 国际安全极客大赛中，有选手用对抗样本攻击亚马逊名人识别系统，让主持人蒋昌建的照片被识别为施瓦辛格。同样，掌握了对抗样本（譬如纸条上的图案），只需在脑门上贴上纸条，就可以“戳瞎”系统，破解身份认证的唯一性。

对抗样本攻击凸显了人脸识别技术的脆弱性，对安全攻防提出了新要求。谷歌于 2017 年举办的对抗样本攻防赛旨在加快研究对抗样本，提升机器学习的鲁棒性。俗话说，魔高一尺道高一丈，技术引发的系统羸弱，最终也须技术予以修复。

安全攻防表面上是一场算法间的较量，从产业的角度出发，其所体现的是技术滥用（包括造假、欺骗等）的“低成本、高利润”特征。

在成本方面，首先，开源社区虽然显著降低了 AI 应用开发门槛，与此同时也降低了攻击算法的获取成本。在换脸“奥巴马”案例中，DeepFake 技术在 Github 上开源后，非专业用户也可轻松凭借一张照片生成伪造视频。譬如，2019 年初，就有网友利用开源算法将朱茵版《射雕英雄传》黄蓉的脸换成杨幂，上传到 B 站上，引发网友热议。

其次，相对快速、花式的算法创新，监管滞后也为技术滥用提供了时间的温床。还以 DeepFake 技术为例，该技术早在 2017 年底出现，法律监管却整整滞后了两年。直至 2019 年初，我国网信办才开始对 DeepFake 展开严格监管。相比慢两拍的监管，DeepFake 技术则在两年内飞速进化，并衍生出海量的恶意应用。

相对低门槛的技术获取，技术滥用所带来的不菲利润，是安全风险和黑灰产的“催化剂”。

譬如上文中提到，互联网合成视频以色情内容为主。据报道称，很多网店不仅提供成品视频，还接受私人订制，只要买家提供 20 张照片，就可以快速生成定制化视频，售价约为 1 分钟 20 元到 50 元不等。在某些 QQ 群的黑产交易中，利用动态人脸识别攻击技术，一个支付类账号破解售价约为 10 元，这样的生意一天能做到 8000 元到 10000 元的流水，而售卖技术教学，一个人学费大概在 4000 元左右。

随着人工智能技术的发展，信息安全受到各国的高度重视。世界各主要国家都在加强网络和信息安全领域的布局和竞争。Gartner 预测，2020 年全球信息安全类支出预计将增长 2.4%，达到 1238 亿美元。其中，中国安全市场支出将增长 7.5%，达到 299 亿人民币。按照信息安全支出占比来看，中国信息安全支出占比低于全球。

攻防是信息安全的本质。信息的复制成本几近于零，这也就诞生了网络经济，催生了互联网安全产品市场，造就了 360 等上规模的网安公司。

AI 算法作为信息技术之一，通过信息复制，同样可以摊薄技术的生产成本。对于愈加复杂的信息网络，闭门造车、自扫门前雪的方式显然是不经济的，引导 AI 对抗技术的商业化落地，是加强网络安全的重要途径之一，也是促进 AI 技术应用和推广的安全保障。

对于 AI 防御方而言——主要为 AI 研发、应用类企业，及评测、监管机构等，除自磨利刃外，可能更好的选择是与安全市场合作共赢，以提升系统鲁棒性，构筑 AI 防御护城河。