浙商证券股份有限公司
ZHESHANG SECURITIES CO.LTD

AI安全系列：以子之矛，攻子之盾

——从deepfakes深度伪造技术看AI安全

行业评级：看好

2023年6月26日

分析师 刘雯蜀
邮箱 liuwenshu03@stocke.com.cn
证书编号 S1230523020002

研究助理 刘静一
邮箱 liujingyi@stocke.com.cn

摘要

1、Deepfakes技术始于2014年，随着AI大模型能力的提升关注度持续增高

Deepfakes定义可分为广义和狭义两个层次，其中广义上指**利用了以生成对抗网络技术（GAN）为主体的深度学习技术制造的看起来很真实但实际上属于虚假的图片或视频。**随着AI大模型的能力不断突破，deepfakes受到的关注度持续增高。

2、 生成式 AI 模型是Deepfakes的技术基础，而人脸伪造技术是deepfakes的一个重要分支

生成式 AI 是深度学习的一个分支，根据生成内容的类别，**生成式AI模型可进一步被分为生成式语言模型和生成式图片模型。**而随着多模态模型的应用逐步深入，**生成式AI模型也开始向多模态方向发展。**目前主流生成式AI模型包括VAE、GAN、diffusion模型等，stable diffusion模型的发布再次使得生成式AI的输出质量大幅提升。

人脸伪造技术是deepfakes的一个重要分支，可被进一步划分为有目标可视身份伪造和无目标可视身份伪造，其中**有目标可视身份伪造技术已经在俄乌冲突中有所应用。**

3、生成式AI技术的迭代增加伪造图片的真实度，同时增大AIGC识别难度

考虑到：1）**数据层面**，可供训练的数据集规模扩大，带来模型效果的大幅提升；2）**算法层面**，算法框架的迭代与组合，使得模型训练更加高效、稳定；3）**算力层面**，算力水平的提升，助力更加高效的复杂模型训练，AI伪造图片的真实度不断增加，使得识别难度增大，带来的潜在风险提升。

4、反AI生成市场空间约64亿元

针对AI生成图像可能带来的风险，各国政府陆续推出政策加以约束。根据我们测算，**反AI生成市场空间约64亿元，其中监管侧44亿元，企业侧20亿元。**

5、部分反AI生成相关公司

上市公司：美亚柏科、东方通
非上市公司：瑞莱智慧、中科睿鉴、Illuminarty（海外）、AI Voice Detector（海外）、Google（海外）

风险提示

1、AIGC监管以及识别技术研发以及相关产品落地不及预期

2、报告中对各类模型的介绍与总结基于对于相关论文内容的理解，具有一定主观性

3、报告中对于反AI生成的市场空间测算存在主观判断及口径差异

4、由AI安全需求带来的市场竞争加剧

5、板块政策发生重大变化

目录

CONTENTS

01 Deepfakes 演进与发展历程

1）Deepfakes：基于深度学习的AI换脸技术

Deepfakes定义可分为**广义和狭义**两个层次：

- **广义：**Deepfakes指使用机器学习（ML）技术创建的伪造品，即利用了**以生成对抗网络技术（GAN）为主体的深度学习技术**制造的看起来很真实但实际上属于虚假的图片或视频。
- **狭义：**Deepfakes是“Deep machine learning”和 “fake”的组合词，是一种**基于深度学习的人物图像合成技术。**不同于一般意义上的p图，Deepfakes利用电脑程序找到两个面部之间的共同点，通过搭建神经网络来学习人脸，使替换以后的脸可以生动地模仿原来的表情，以假乱真。

2）Deepfakes的产生与飞跃

- **2014年是Deepfake的诞生元年**，Goodfellow与同事发表的科学论文标志着GAN AI的诞生，催生出我们如今所熟知的Deepfakes。在2014年，就有迹象表明 GAN 有望生成仿真度极高的人脸。深度伪造技术开始进入大众视野。在Deepfakes产生初期，生成代理往往倾向于产出分辨率较低而模糊不清的图像是检查代理难以判断内容的真伪，深度伪造技术一定阶段内存在着**输出内容像素过低，生成结果难以令人信服**的问题。
- **2017年英伟达推动质量飞跃**， 利用分阶段训练网络解决了生成代理产出分辨率过低的问题，GAN 开始产出质量空前的伪造人像，**深度伪造技术开始被推向市场主流。**自此，**Deepfakes一词成为 AI 生成图像和视频的代名词。**
- **2019年Deepfake正式成为市场主流，**专注于Deepfakes的YouTUbe频道拥有数百万关注者，产出质量远高于其他AI模型。

资料来源：AI前线公众号、浙商证券研究所

01 Deepfakes技术发展历程

- **2014：**Goodfellow与同事发表了全球首篇介绍GAN的科学论文，代表着**GAN AI**的诞生，催生出如今为人熟知的 Deepfakes。
- **2015：**研究人员开始将 **GAN 与经过图像识别优化的多层卷积神经网络（CNN）相结合**，这一组合取代了以往较为简单的 GAN 代理驱动网络，提高了处理数据的速度和显卡运行效率，也让生成结果的可信度迈上新的台阶。
- **2016：**研究人员把两个 GAN 结合起来，开展**并行学习。**
- **2017：**英伟达推动质量飞跃， GAN 开始产出质量空前的伪造人像，**深度伪造技术开始被推向市场主流。**自此，**Deepfakes一词成为了 AI 生成图像和视频的代名词。**
- **2018：**英伟达提升 GAN 控制能力，使其能够对人像中的“黑发”和“微笑”等图像单一特征作出调整，**将训练图像中的特征有针对性地转移到 AI 生成图像当中。**
- **2019：**三星公司的研究人员公布了一种能够深度伪造人类和艺术品的 GAN，只需参考少数照片就能利用Deepfakes AI达成出色的伪造效果；以色列研究人员又推出了**换脸 GAN（FSGAN）**，能够对即时视频中的人脸进行实时交换。无需任何预先训练。
- **2020：**微软推出 FaceShifter，该软件能够利用模糊的原始图片，依赖于分别负责**伪造人脸和照片比对**的两套网络，生成高度可信的 Deepfakes 图像；深度伪造技术有望成为迪士尼电影制作开发的主流技术。
- **2021：**社交媒体中出现Deepfakes**巡演、直播与人脸租赁**活动，在市场上获得极高热度。
- **2022：**GAN的改进接连出现，包括能够在短视频片段中轻易操纵人脸的**StyleGAN2变体**和既能以高度匹配的 3D 形式生成统一图像，也能利用一张真人图像还原出 3D 模型的**3D GAN**，大力推动力AI深度伪造技术的发展。

资料来源：AI前线公众号、浙商证券研究所

01 Deepfakes应用发展历程

1）海外应用

- **FaceShifter：**2020年由北京大学和微软亚洲研究院研究团队联合发表，是一种**高保真、能够感知遮挡**的AI换脸工具，采用两层框架结构实现高精度和遮挡条件下的换脸。其优于以往同类技术，在生成逼真的人脸图像方面表现优异，被誉为机器学习图像识别领域的“利矛”。
- **Wombo AI：**2021年正式进入大众视野，可以借助AI技术**将声音与图片中的角色自动对上口型**，使处于静止图片中的人物进行开口讲话，并且还有会动的姿态表情，在社交媒体上大受欢迎。
- **DeepFaceLive：**由DeepFaceLab的缔造者在 2021 年首次展示，能够在经过适当训练、或者接收到预训练 AI 模型之后，在**实时视频中交换人脸**，意味着换脸的技术又再一次的突破。

2）国内应用

- **Zao：**于2019年首次公测，是中国国内利用**深度伪造技术**制作的一个应用程序，用户可以利用这个程序将自己的脸替换成电影里某个角色的脸；用户还可以在Zao里面大量的视频和图片库中进行选择，在上传视频后就可以在几分钟之内生成深度伪造的角色。然而，从2019年9月1日，Zao就由于疑**侵犯用户肖像权**而被用户投诉，以及**对用户生物识别信息的采集存在的信息安全性问题带来的安全风险**而遭遇下架。
- **Face X-Ray：**2020年由北京大学和微软亚洲研究院研究团队联合发表，是一种针对**伪造人脸图像**的通用检测工具，不需要依赖于与特定人脸操作技术相关的伪影知识，并且支持它的算法可以在不使用任何方法生成假图像的情况下进行训练。这种工具能有效地识别出未被发现的假图像，并能可靠地预测混合区域，在市场上获得很高的评价，被誉为机器学习图像识别领域的“坚盾”。

资料来源：新智元公众号、《FaceShifter: Towards High Fidelity And Occlusion Aware Face Swapping》、《 Face X-ray for More General Face Forgery Detection》、品牌观察报公众号、浙商证券研究所

Deepfakes应用

表：国内外deepfakes相关应用梳理

名称	类别（开源与否）	开发地区	开发时间	使用状况
FakeApp		国外	2018	已下架
Faceswap	已开源	国外	2019	未下架
DeepFaceLab	已开源	国外	2018	未下架
DeepFaceLive	已开源	国外	2019	未下架
Faceswap-GAN		国外	2019	未下架
ZAO	已开源	国内	2019	已下架
DFaker		国外	2020	未下架
Deepface		国外	2015	未下架
FaceShifter		国外	2020	未下架
Wombo AI		国外	2021	未下架
Avatarify	已开源	国外	2020	未下架

资料来源：《深度伪造人脸视频检测技术综述》、Faceswap官网、DeepFaceLab官网、DeepFaceLive官网、搜狐网、重蔚自留地、浙商证券研究所

02 生成式AI模型梳理