深度学习视域下生成式伪造手写数字鉴定方法

张梦萍[1] 牟 熠[1] 曾 敏[1] 肖尧馨[1]
ZHANG Mengping MOU Yi ZENG Min XIAO Yaoxin

摘 要 随着基于GAN模型的生成式技术的发展，传统伪造手写数字鉴定方法面临着巨大挑战。为此，采用多种GAN模型生成手写数字，并构建相关数据集后，利用Vision Transformer模型实现了伪造手写数字鉴定，并可以进行溯源。通过构建工程软件并进行实验测试，实验结果表明系统在伪造真假鉴定和伪造手写数字来源两方面的平均准确率、精确率、召回率分别达到79.16%和77.22%、80%和81.52%、97.22%和91.24%，达到了预期的实验目标，为识别生成式伪造手写数字提供了可参考的鉴定方法。

关键词 GAN；手写数字；伪造鉴定；Vision Transformer

doi：10.3969/j.issn.1672-9528.2024.08.046

0 引言

生成式人工智能技术是当前深度学习领域最热门的研究方向之一，其主要利用生成对抗网络（generative adversarial network，GAN）模型，实现各种音视频、文本等信息的生成、修复与合成。随着社交媒体传播信息的加快，以及生成式模型工具使用门槛的逐步降低，其带来内容生成的爆发性增长的同时，也导致了潜在的安全风险。深度伪造，是由“deep learning”（深度学习）和“fake”（伪造）组合起来的词，意思是通过深度学习算法，实现音视频的生成和伪造[1]。简言之，“Deepfake”是一种利用深度学习算法来篡改图形图像的技术[2-3]。人工智能技术尤其是Deepfake技术的滥用，使得社会公民无法辨别出真正的信息，从而引发广泛的信任危机。因此，很有必要对深度伪造图像进行检测，识别出伪造图像，并检测出伪造的具体信息。国内关于深度伪造检测技术也进行了较多综述。王任颖等人[4]对视听觉的深度伪造检测技术进行了综述。李旭嵘等人[5]对深度伪造生成技术和深度伪造图像检测技术进行了综述。李颖等人[6]基于CNN（convolutional neural networks）与Tansformer模型实现了对伪造视频的高效检测。石达等人[7]基于CycleGAN实现了伪造性别图像生成。张小娜[8]开发了一套基于Java语言的伪造图像检测系统。这些已有的方法模型都基于通用伪造数据集而训练，如Face Forensics++[9]、CelebA，但现有的生成式模型众多，每种模型生成的伪造图像各具特点，使得现有方法模型受到挑战，且现有方法还只能进行图像是否伪造的判定，而很难溯源图像伪造的过程细节，进而使得其在法律层面作为证据时的充分性略显不足。据此，本文以众多GAN网络生成的手写数字图片为数据集，构建了以Vision Tansformer为基础的生成式伪造数字鉴定模型，其在判定手写数字是否伪造的同时，也可推测伪造数字图片的算法来源。

1 生成对抗网络模型原理及其生成数据构成数据集

生成对抗网络主要包括判决器和生成器两个模块，在其相互之间的对抗训练中，逐步提升生成器的生成能力与判决器的判定能力。基于深度学习的生成对抗网络最早由Ian Goodfellow提出，主体结构如图1所示，其利用式（1）构成损失函数。

$$\min_G \max_D V(D,G)=\mathsf{E}_{x\sim p_{data}(x)}[\log D(\boldsymbol{x})]+\mathsf{E}_{z\sim p_z(z)}[\log(1-D(G(\boldsymbol{z})))] \quad (1)$$

式中：x~Pdata(x)表示数据来源于真实样本。$D(x)$为真实图片的概率。z~pz(z)表示数据来源于随机噪声。$D(G(Z))$表示噪声z通过生成器生成图片为真实图片的概率。

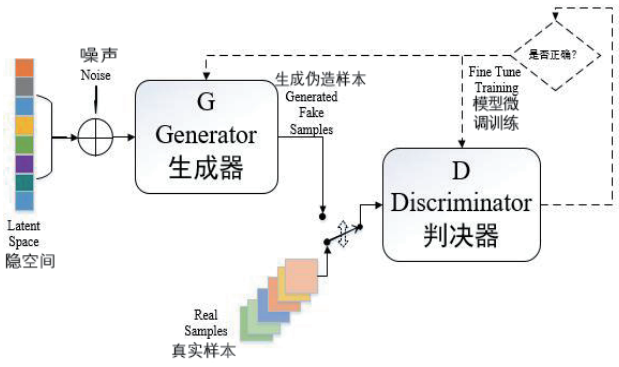

图1 生成对抗网络结构图

1. 内江师范学院人工智能学院 四川内江 641100

[基金项目]深度伪造图像检测技术（No.2022YB24）；大学生创新创业项目“深度伪造图片检测技术”（No.X202203）；大学生创新创业项目“创界守护——防止AI盗取便携式创意加密盒”

随着 CNN 在图像处理领域表现出更优异的性能，Radford 等人将 CNN 与 GAN 结合，提出了深度卷积生成对抗网络（deep convolution generative adversarial networks，DCGAN）。其结构如图 2 所示，整个网络模型不使用全连接，而是直接基于卷积与反卷积运算，最终实现二维图像的生成。

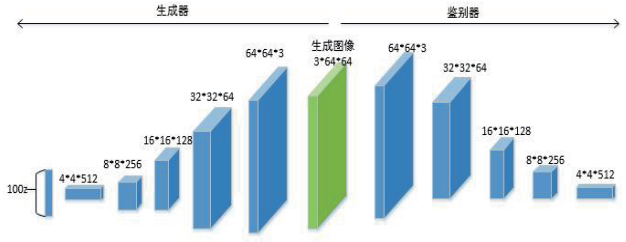

图 2 DCGAN 结构图

由于在训练原始 GAN 网络时，输入生成器的数据是一组随机信号，具有严重的不可控性，其真实的训练效果相对预期可能产生严重的偏差，为此，Mehdi Mirza 等人提出了 CGAN（conditional generative adversarial networks），即条件对抗生成网络。该网络将真实数据与标签共同作为输入来训练模型，从而提升了模型的方向性与有效性，对应结构如图 3 所示，损失函数如式 2 所示 [10]。

$$\min_G \max_D V(D,G)=\mathsf{E}_{x\sim p_{data}(x)}[\log D(\boldsymbol{x}\mid \boldsymbol{y})]+\mathsf{E}_{z\sim p_z(z)}[\log(1-D(G(\boldsymbol{z}\mid \boldsymbol{y})))] \quad (2)$$

式中：log$D(x|y)$ 表示给定条件下判别器对生成器生成的样本的预测为真的对数概率。log(1-$D(G(z|y))$ 指给定条件 y 下判别器 D 对生成器 G 生成的样本 $G(z|y)$ 的预测为假的对数概率。

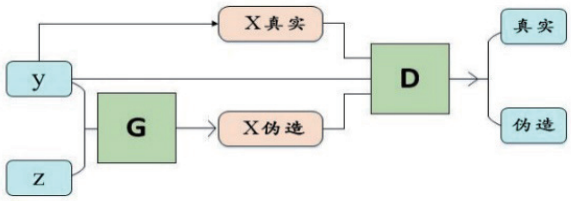

图 3 CGAN 结构图

编解码器是深度学习中的重要模型之一，Anders Boesen 等人将其与 GAN 结合，提出了 VAE-GAN 模型。VAE 是 Variational Auto-Encoder 的缩写，含义是变分自编码器。该模型的前后两部分共用一个结构，其在前部分的编解码中充当解码器，在后半部分中充当生成器。VAE 利用 GAN 的判决能力降低解码图像的模糊问题，GAN 利用 VAE 获得更可控的生成器，避免随机噪声输入生成器而造成模型训练瓶颈问题 [11]。

本文搭建 Ian Goodfellow 提出的原始 GAN、Radford 等人改进形成的 DCGAN、Mehdi Mirza 等人改进提出的 CGAN、Anders Boesen 等人改进提出的 VAE-GAN 等模型，并利用 MINIST 数据集，实现每种模型手写数字图片的生成。各个模型生成典型结果如图 4 所示，最终每个模型对每个数字生成 7000 张有效图片，并标记其来源生成对抗网络类型。结合 MINIST 数据集已有数据，使得本文构建的数据集包含 350k 张图片。所生成的数字图片如图 5 所示。

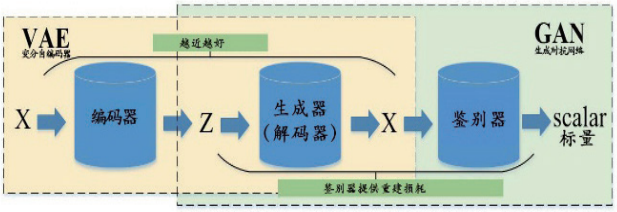

图 4 VAE-GAN 模型结构图

图 5 手写数字图片生成

2 本文方法原理

本文基于 Vision Transformer 模型实现生成式伪造数字图片的鉴定，该模型结构如图 6 所示。Transformer 是自然语言处理领域的重要模型之一，通过使用多头注意力机制，极大提升了各场景的处理效果，A Dosovitskiy 等人以其为基础，将输入图像进行分块而形成序列图像来模拟自然语言中语句单词的前后连接关系，进而实现了图像分类效果的提升。该模型损失函数如式（1）所示。

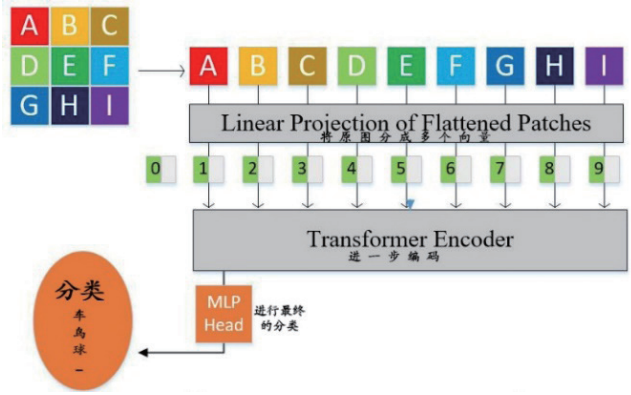

图 6 ViT 模型图

本文的数据集中包括 5 种类别的手写数字，因此 ViT 模

型的最终输出分类也设定为 5 个类别，由此本文构建了如图 7 所示的生成式伪造数字图片鉴定框架。鉴定框架将输入手写数字图像缩放至 224×224 大小，并进行 16×16 的分块，在输入 ViT 模型后实现对输入手写数字图像类别鉴定。

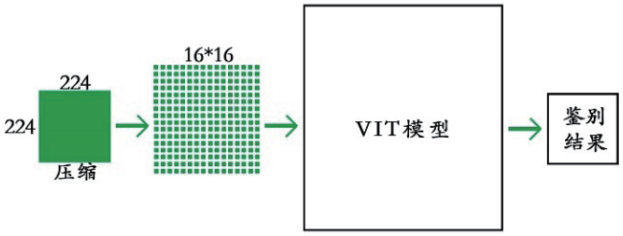

图 7 生成式伪造数字图片鉴定框架

3 实验

3.1 实验环境及评价指标

模型训练及实验测试时，硬件参数主要为 Intel Core i9-9900K CPU、NVIDIA A100 GPU、64 GB 内存，软件环境主要包括 PyTorch 1.7、OPECV、Python。数据集使用时，按照 6:2:2 的比例划分为训练集、验证集、测试集。模型训练时，学习率设置为 0.000 01，学习率每 10epoch 衰减 1 次，批次设置为 32 个和 40 个 epoch 完成模型训练。

为对实验结果进行量化评估，主要针对伪造真假鉴定 J1 和伪造手写数字来源 J2 等 2 个指标，计算准确率（accuracy）、精确率（precision）、召回率（recall）。准确率定义如式（3）所示，表示给定的数据，分类正确的样本数占总样本数的比例。其中，TP 表示伪造样本检测正确，TN 表示伪造样本检测错误，FN 表示非伪造样本检测正确，FP 表示非伪造样本检测错误。

$$ACC=\frac{TP+TN}{TP+FN+FP+TN} \quad (3)$$

精确率定义如式（4）所示，用于表达模型判别为伪造数字中实际正确的概率。Precision 相对正例的预测结果而言，正例预测的准确度。

$$Precision=\frac{TP}{TP+FP} \quad (4)$$

召回率定义如式（5）所示，用于表达真实伪造数字中被判断出来为正例的概率。Recall 以实际样本为判断依据，实际为正例的样本中，被预测正确的正例占总实际正例样本的比例，评估所有实际正例是否被预测出来的覆盖率占比多少。

$$Recall=\frac{TP}{TP+FN} \quad (5)$$

3.2 实验结果与讨论

实验过程开发的软件系统如图 8 所示，主要包括 4 种 GAN 模型伪造手写数字生成过程、ViT 判定伪造手写数字过程及来源追溯。

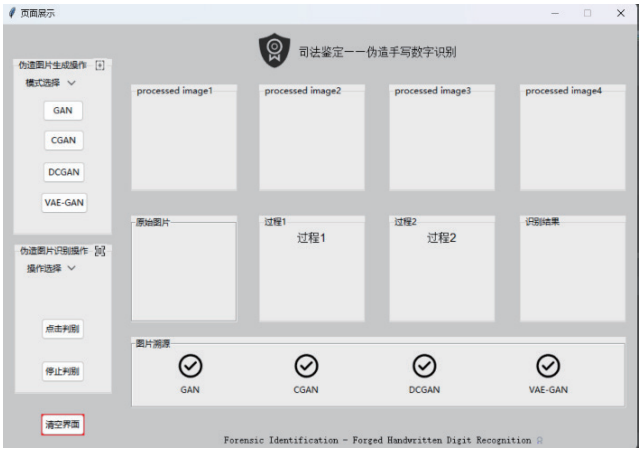

图 8 软件系统界面

对于指标 J1，其实验量化结果如表 1 所示。

表 1 JI 指标实验量化结果

数据来源	GAN	DCGAN	CGAN	VAE-GAN	MINIST	平均值
Accuracy	79.6	76.7	75.9	78.2	85.4	79.16
Precision	80.1	76.8	77.8	78.8	86.5	80
Recall	97.5	97.7	95.5	97.6	97.8	97.22

对于指标 J2，其实验量化结果如表 2 所示。

表 2 J2 指标实验量化结果

数据来源	GAN	DCGAN	CGAN	VAE-GAN	MINIST	平均值
Accuracy	72.7	72.2	82.7	74.5	84	77.22
Precision	77.1	78.7	86.4	78.7	86.7	81.52
Recall	90.2	88.1	92.7	90.1	95.1	91.24

4 结论

本文基于深度学习实现了生成式伪造手写数字图片的鉴定，且可识别产生伪造手写数字图像的生成式模型。为此，首先利用多个生成对抗网络模型生成众多手写数字图像，并结合 MINIST 数据集构建伪造手写数字数据集，随后以该数据集训练 VIT 网络模型，从而实现生成式伪造手写数字图片的识别与分类。由于实验所采用的数据集仅来自 4 种 GAN 模型，具有一定的局限性，后期的研究中将逐步扩大数据生成来源，以增加 VIT 网络鉴定能力。

参考文献：

[1] 李斌 . 基于深度学习的图像内容辨别方法研究 [D]. 西安 : 西安理工大学 ,2021.

[2] 谭维瑾 . 基于深度神经网络的图像伪造定位和检测算法研究 [D]. 南京 : 南京信息工程大学 ,2021.

（下转第 206 页）

因此，在推进多模态数据融合技术在新闻采编中的应用时，需要综合考虑技术、法律、伦理等多方面的因素，确保技术的合规性和可持续性。同时，也需要加强技术研发和人才培养，为新闻媒体的数字化转型和创新发展提供有力支撑。

参考文献：

[1] 魏丹阳 . 生成式人工智能在新闻采编工作中的应用与挑战 [J]. 新闻文化建设 ,2024(3):94-96.

[2] 高慧琳 . 基于麦克卢汉媒介观的新媒介技术哲学研究 [D]. 大连 : 大连理工大学 ,2022.

[3] 刘晓倩 , 张英俊 , 秦家虎 , 等 . 模糊认知图学习算法及应用综述 [J]. 自动化学报 ,2024,50(3):450-474.

[4] 徐琦 , 赵子忠 . 中国智能媒体生态结构、应用创新与关键趋势 [J]. 新闻与写作 ,2020(8): 51-58.

[5] 王宝莹 . 新媒体视域下的高校舆情传播研究 [D]. 哈尔滨 : 黑龙江大学 ,2024.

[6] 郑文锋 . 全媒体传播时代新闻策划理念变迁与创新路径 [J]. 传播与版权 ,2024(10):1-4.

[7] 许娜 . 人工智能技术在新闻制作过程的应用 [J]. 电视技术 , 2024, 48(3):103-105.

[8] 闵媛春 , 杨明义 . 我国 AI 主播研究的可视化图谱分析 [J]. 传媒论坛 ,2024,7(9):71-74.

[9] 孙立峰 , 宋新航 , 蒋树强 , 等 . 多模态协同感知与融合技术专题前言 [J]. 软件学报 ,2024,35(5):2099-2100.

[10] 段毛毛 , 魏燚伟 . 基于多模态交互网络的图像描述 [J]. 计算机技术与发展 2024,34(5):44-51.

[11] 刁生富 , 陈惠 . 元宇宙背景下虚拟数字人对人类的影响探究 [J]. 佛山科学技术学院学报 (社会科学版) , 2024, 42(3): 31-37.

[12] 胡新华 . 虚拟数字人技术在新闻传播中的应用研究 [J]. 科技传播 , 2024, 16(3):1-3+7.

【作者简介】

李满江（1970—），男，河北蠡县人，本科，高级工程师，研究方向：中文信息处理。

鞠传森（1981—），男，山东济南人，硕士研究生，正高级工程师，研究方向：大数据应用、媒体融合。

任鹏（1971—），男，山东高密人，本科，工程师，研究方向：计算机应用。

（收稿日期：2024-06-14）

（上接第 201 页）

[3] 邢豪 . 基于时空特征的深度伪造视频篡改检测 [D]. 太原 : 太原理工大学 ,2021.

[4] 王任颖 , 储贝林 , 杨震 , 等 . 视觉深度伪造检测技术综述 [J]. 中国图象图形学报 ,2022(1):43-62.

[5] 李旭嵘 , 纪守领 , 吴春明 , 等 . 深度伪造与检测技术综述 [J]. 软件学报 , 2021,32(2):496-518.

[6] 李颖 , 边山 , 王春桃 , 等 .CNN 结合 Transformer 的深度伪造高效检测 [J]. 中国图象图形学报 ,2023,28(3):804-819.

[7] 石达 , 芦天亮 , 杜彦辉 , 等 . 基于改进 CycleGAN 的人脸性别伪造图像生成模型 [J]. 计算机科学 ,2022,49(2):31-39.

[8] 张小娜 . 基于 Java 语言的伪造图像识别检测算法 [J]. 单片机与嵌入式系统应用 ,2021,21(10):49-53.

[9]RÖSSLER A, COZZOLINO D, VERDOLIVA L, et al. Face forensics ++: learning to detect manipulated facial images[C]// Proceedings of 2019 IEEE / CVF International Conferenceon Computer Vision (ICCV). Piscataway: IEEE, 2019: 1-11.

[10]Conditional Generative Adversarial Nets（CGAN）[EB/OL].(2020-07-20)[2024-01-25].https://blog.csdn.net/qq_40128284/article/details/107458339.

[11] Keep_Trying_Go. WGAN 基本原理及 Pytorch 实现 WGAN [EB/OL]. (2023-05-03)[2024-01-26].https://blog.csdn.net/Keep_Trying_Go/article/details/130471766.

【作者简介】

张梦萍（1989—），女，四川宜宾人，硕士研究生，助教，研究方向：计算机应用、教育技术学、现代远程教育。

牟熠（2004—），女，四川宜宾人，本科在读，研究方向：计算机科学与技术。

曾敏（2004—），女，四川内江人，本科在读，研究方向：计算机科学与技术。

肖尧馨（2004—），女，四川资阳人，本科在读，研究方向：计算机科学与技术。

（收稿日期：2024-06-12）