基于深度学习的多特征融合人脸鉴伪模型

Multi-feature Fusion Face Authentication Model Based on Deep Learning

李　铮[1],郑　涛[2],张小梅[2](1. 中讯邮电咨询设计院有限公司,北京 100048;2. 中国联合网络通信集团有限公司,北京 100033)
Li Zheng[1],Zheng Tao[2],Zhang Xiaomei[2](1. China Information Technology Designing & Consulting Institute Co.,Ltd.,Beijing 100048,China;2. China United Network Communications Group Co.,Ltd.,Beijing 100033,China)

摘　要:
人脸伪造给网络安全带来了重大挑战。针对现有人脸鉴伪模型特征单一、准确率低的问题,提出了一种基于深度学习的多特征融合人脸鉴伪模型。该模型设计了不同特征提取模块,用以获取不同尺度的特征表示。并学习如何有效融合这些语义信息以准确判定是否伪造,从而显著提升模型的准确率和鲁棒性。最后在公开数据集FaceForensics++上进行大量实验验证。实验结果显示,与现有方法相比,设计的模型有明显的性能提升。

关键词:
人脸鉴伪;特征融合;深度学习;FaceForensics++数据集
doi:10.12045/j.issn.1007-3043.2024.08.012
文章编号:1007-3043(2024)08-0058-04
中图分类号:TP391
文献标识码:A
开放科学(资源服务)标识码(OSID):

Abstract:
Face forgery poses a major challenge to network security. In response to the problem of existing face forgery models with single features and low accuracy, a multi-feature fusion face forgery model based on deep learning is proposed. The model designs different feature extraction modules to obtain feature representations at different scales. It also learns how to effectively fuse this semantic information to accurately determine whether it is forged, thereby significantly improving the accuracy and robustness of the model. Finally, a large number of experiments are carried out on the open data set FaceForensics++. The experimental results show that the designed model achieves significant performance improvement compared to existing methods.

Keywords:
Face forgery detection;Feature fusion;Deep learning;FaceForensics++ dataset

引用格式:李铮,郑涛,张小梅. 基于深度学习的多特征融合人脸鉴伪模型[J]. 邮电设计技术,2024(8):58-61.

1 概述

随着生成对抗网络(Generative Adversarial Networks,GAN)的快速发展,人脸深度伪造技术取得了巨大的进步,生成的图像逼真,令人难以分辨。为了区分真伪,相关科研人员努力探索各种检测人脸伪造的方法[1-5]。早期相关技术主要依赖于手工提取各种特征进行伪造检测[6-7],导致模型准确率低且泛化能力和鲁棒性受到了限制。因此,基于深度学习(Deep Learning,DL)的相关人脸鉴伪方法成为了近年来的主要研究方向。同时这个领域也涌现出了一系列的人脸鉴伪方法。例如Afchar等人[8]提出的一种紧凑型面部视频伪造检测网络,通过提取一组有效的面部微表情、质地等特征,从而区分真实视频和伪造视频。Liu等人[9]提出了一种全局纹理增强方法,用于野外环境下的人脸检测,通过增强图像的全局纹理信息,提高了人脸检测系统对伪造人脸的敏感性,从而提升检测的准确性和鲁棒性。Chen等人[10]提出了一种将频域信息与RGB信息相结合的检测方法,通过离散余弦变换过滤低频特征并保留高频特征,同时结合图像中提取的低层、中间层和高层卷积特征,将2种不同模态的信息进行融合,从而提高模型的鲁棒性。杨挺等人[11]提

收稿日期:2024-06-14

出了基于改进三元组损失的伪造人脸视频检测方法，通过改进传统的三元组损失函数，更好地引导模型学习到有效的特征表示，从而提高检测效果。Coccomini等人[12]提出了一种基于Transformer的伪造检测方法，将视觉Transformer与卷积特征提取器结合起来，利用EfficientNet B0作为特征提取器与Vision Transformers相结合，从而实现了对视频中深度伪造的检测。尽管目前的研究算法已经在人脸鉴伪方向取得了一些成绩，但仍然存在一些挑战。为了克服这些挑战，并进一步提高人脸鉴伪系统的性能，本文提出了一种基于深度学习的多特征融合人脸鉴伪模型，旨在克服传统方法中存在的局限性，并提高人脸鉴伪系统的性能。该模型通过融合多种特征信息，包括图像外观、纹理等方面的特征，构建了一个更加全面和准确的人脸鉴伪模型。与单一特征的方法相比，多特征融合能够提供更多的信息，增强了系统对真实和伪造人脸的区分能力。

2 基于深度学习的多特征融合人脸鉴伪模型

2.1 模型整体结构

本文为了提升伪造人脸视频检测模型的准确率，提出了一种深度伪造高效转换网络（Deepfake Efficient-Feature Transformation Network，DEFTNet），该网络的整体架构包括主干网络模块、宽接受域局部特征提取模块与混合特征融合模块（见图1）。

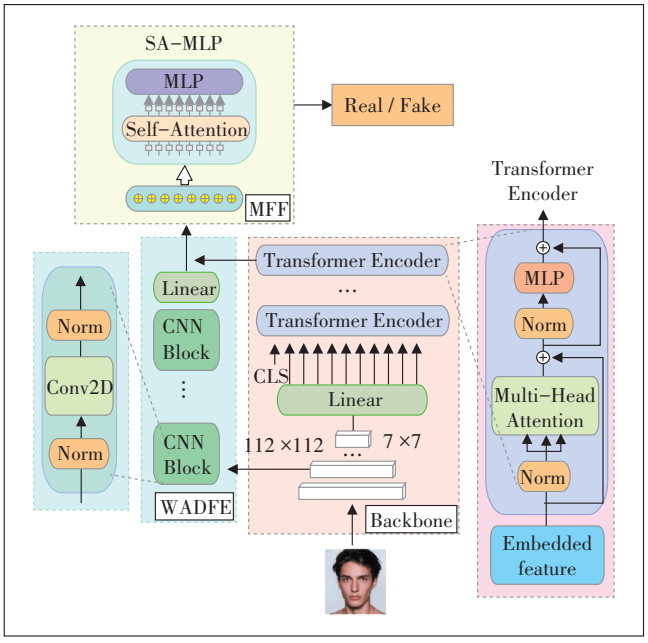

图1　模型框架

模型输入为人脸二维数据。第1步是由EfficientNet网络[13]和transformer encoder组成[14]的主干网络部分，主要用于小窗表征提取阶段，第2步为获取EfficientNet中间结果进行宽接受域局部特征提取。第3步主要作用是特征融合，将特征提取部分的2个阶段结果进行融合，从而进一步获取到真假样本特征的高纬度表示。

2.2 主干网络模块

主干网络使用高效卷积网络提取图像的多通道高维卷积特征，并通过视觉编解码器（Vision Transformer，ViT）架构[15]进行处理。高效卷积网络基于EfficientNet B4卷积网络提取人脸patch小窗表征，大小为7×7。EfficientNet B4卷积网络采用了预训练的权重，获取到网络的最后一层输出，通过使用EfficientNet B4卷积网络提取的特征，简化ViT的训练过程。EfficientNet B4是EfficientNet系列中的一个具体变种，其网络结构由一系列重复的卷积层、批归一化层和激活函数组成。这些模块的数量和宽度会根据网络的深度和宽度进行调整，以保持网络的效率和性能。

通过EfficientNet卷积网络得到的局部特征转换成序列形式输入到Transformer encoder使其获取到全局范围内的图像语义特征，Transformer encoder通过自注意力机制、多头注意力机制、残差连接和层归一化这些关键组件的结合，能够有效地捕捉输入序列的长期依赖关系，从而捕捉全局的图像语义特征。

2.3 宽接受域局部特征提取模块

该模块使用EfficientNet网络结构中reduction层输出表征作为输入，考虑到第1阶段EfficientNet卷积网络得到的特征是小窗表征，可能并不是一个很完备的选择，所以使用reduction_1层进行大窗口表征的提取以获得更宽的接受域局部特征，大窗表征大小为112×112。WADFE是一种深度学习网络，通过多层卷积网络和非线性的组合，实现对输入数据进行特征提取和变换（见图2）。

基于深度学习的核心思想是通过多层非线性变换，逐渐提取输入数据的高层抽象特征。在卷积层之后添加非线性激活函数，如ReLU函数，使其能够学习到更复杂的特征表示。通过池化层，对特征图进行降采样，减少特征图的维度，提高模型的计算效率并增强模型的平移不变性。在经过多层卷积和池化之后，利用全连接层对特征图进行展平并进行线性变换得到全局的特征表示。

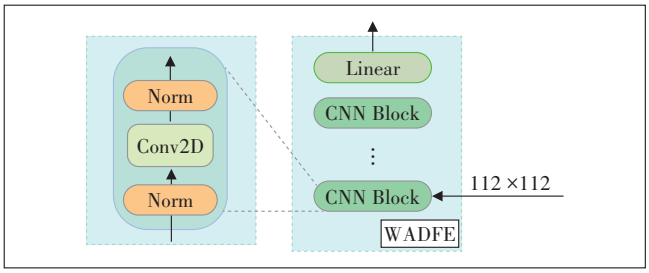

图2　WADFE框架

$$\mathrm{ReLU}(x)=\begin{cases}x, & x\geqslant 0\\ 0, & x<0\end{cases}=\max(0,\ x) \qquad (1)$$

2.4 混合特征融合模块

特征融合是将关键信息整合的过程,旨在将来自多个特征源的信息结合在一起,以提取出更加全面和丰富的特征表示。

本文采用通道空间级融合策略,将来自不同通道的特征向量进行通道融合,可以充分利用不同通道的信息,提高特征的表征能力,其结构如图3所示。同时将相邻位置的特征向量进行整合,以提取出更具有局部结构信息的特征表示。通过引入自注意力机制,根据不同特征源的重要性加权融合,从而实现对特征信息的自适应整合。通过自注意力机制,模型能够自适应地关注到图像中最相关和最有代表性的区域,从而提高分类的准确性和鲁棒性。

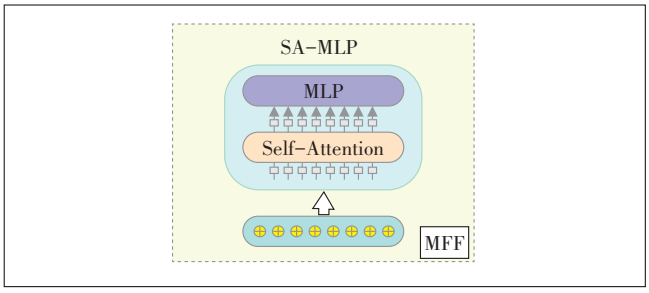

图3　SA-MLP结构

3 实验结果与分析

3.1 数据集与评价指标

3.1.1 数据集

本文所有的实验研究均在公开数据集FaceForensics++[16]的3种不同合成方法上进行。数据集中真实的视频大部分来自YouTube的视频片段,总共包含1 000个视频。伪造的数据分别来自DeepFakes、Faceswap和NeuralTextures 3种伪造方法,每种伪造方法根据真实数据生成对应的1 000个视频。根据数据集的数量将总数据集按7:2:1的比例划分为训练数据集、验证数据集和测试数据集。每个视频抽取38帧数据,因此训练数据集包含人脸数据106 400张,验证数据集包含人脸数据30 400张,测试数据集包含人脸数据15 200张,整个数据集分布如表1所示。

表1　伪造人脸训练数据集分布(单位:张)

数据类型	数据集		
	Train	Val	Test
YouTube	26 600	7 600	3 800
DeepFakes	26 600	7 600	3 800
Faceswap	26 600	7 600	3 800
NeuralTextures	26 600	7 600	3 800
总计	106 400	30 400	15 200

3.1.2 评价指标

本文采用准确率(Accuray,Acc)作为评价指标对实验的结果进行评估,该公式的定义为:

$$\mathrm{Acc}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}\times\frac{1}{2} \qquad (2)$$

其中TP表示真正例(True Positives),即模型正确预测为正类别的样本数量。TN表示真负例(True Negatives),即模型正确预测为负类别的样本数量。FP表示假正例(False Positives),即模型错误预测为正类别的样本数量。FN表示假负例(False Negatives),即模型错误预测为负类别的样本数量。

3.2 实验环境

本文实验装置是一台CPU配置为Intel(R) Xeon(R) Gold 6248R CPU @ 3.00GHz,显卡配置为英伟达的GeForce RTX 3090,内存为256G的服务器。操作系统为Centos 7.6,显卡驱动版本为460.27.04,CUDA版本为11.2。使用Python 3.7.16作为开发语言,Pytorch 1.9.0作为深度学习框架。试验过程中,将人脸缩放到224×224的大小,Batch-Size设置为32,采用2张GPU卡对其进行训练,采用自适应矩估计算法(Adaptive Moment Estimation,Adam)作为模型的优化器,初始化学习率设置为0.1,动量设置为0.9,权重衰减设置为0.000 5,因为采用了多张显卡进行训练,为了保持批量归一化参数计算的一致性,使用同步批量归一化技术对模型进行批量归一化操作,提高模型的性能,改善训练效果。

3.3 消融实验

本文提出的方案是基于文献[12]进行的改进,并与之在公开数据集FaceForensics++上进行测试对比,

实验结果如表2所示。以文献[12]中提出的Efficient ViT为基线（Baseline），在此基础上引入特征融合模块，从结果分析可知，结合特征融合模块的DEFTNet（MLP）模型，DeepFakes提升了2.5个百分点，Faceswap提升了1.34个百分点，NeuralTextures升了0.95个百分点，整体平均准确率提升了1.6个百分点，这表明了特征融合模块的有效性。结合自注意力机制（SA-MLP）之后，在原有的基础上DeepFakes、Faceswap和NeuralTextures又分别提升了0.28、0.77和0.13个百分点，除此之外，将高效网络替换成B4版本，DeepFakes、Faceswap和NeuralTextures又分别提升了0.85、0.75和0.23个百分点，表明了本文方法的效果优异。

表2　消融实验结果（单位：%）

Model	DeepFakes	Faceswap	NeuralTextures
Efficient ViT	83.00	78.00	68.00
DEFTNet（MLP）	85.50	79.34	68.95
DEFTNet（SA-MLP）	85.78	80.11	69.08
DEFTNet（B4）	86.63	80.86	69.31

4 结束语

本文提出了一种基于深度学习的多特征融合人脸鉴伪模型，该模型通过不同感受野获取到不同大小的窗口表征，然后利用人脸的局部特征和全局特征以特征融合的方式将人脸数据进行真伪的辨别。最后通过实验结果分析，该模型能够较好地提升人脸鉴伪的准确率，更好地应对不同人脸伪造算法。下一步，要在现有的模型结构上，在保证准确率基本不变的情况下提升模型的泛化能力，并将本模型轻量化部署在应用系统上以进一步验证在现网环境下模型的鲁棒能力。

参考文献：

[1] LI L Z, BAO J M, ZHANG T, et al. Face X-ray for more general face forgery detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 5000-5009.

[2] ZHAO T C, XU X, XU M Z, et al. Learning to recognize patch-wise consistency for deepfake detection[EB/OL]. [2024-01-11]. https://arxiv.org/pdf/2012.09311v1.

[3] LUO Y C, ZHANG Y, YAN J C, et al. Generalizing face forgery detection with high-frequency features[C]///2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 16312-16321.

[4] SUN K, LIU H, YE Q X, et al. Domain general face forgery detection by learning to weight[C]///Proceedings of the AAAI Conference on Artificial Intelligence. Washington DC: Association for the Advancement of Artificial Intelligence, 2021, 35(3): 2638-2646.

[5] ASNANI V, YIN X, HASSNER T, et al. Reverse engineering of generative models: inferring model hyperparameters from generated lmages[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(12): 15477-15493.

[6] PAN X Y, ZHANG X, LYU S. Exposing image splicing with inconsistent local noise variances[C]//2012 IEEE International Conference on Computational Photography (ICCP). Piscataway: IEEE, 2012: 1-10.

[7] FRIDRICH J, KODOVSKY J. Rich models for steganalysis of digital images[J]. IEEE Transactions on Information Forensics and Security, 2012, 7(3): 868-882.

[8] AFCHAR D, NOZICK V, YAMAGISHI J, et al. MesoNet: a compact facial video forgery detection network[C]//2018 IEEE International Workshop on Information Forensics and Security (WIFS). Piscataway: IEEE, 2018: 1-7.

[9] LIU Z Z, QI X J, TORR P H S. Global texture enhancement for fake face detection in the wild[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 8057-8066.

[10] CHEN S, YAO T P, CHEN Y, et al. Local relation learning for face forgery detection[C]//Proceedings of the AAAI conference on artificial intelligence. Washington DC: Association for the Advancement of Artificial Intelligence, 2021, 35(2): 1081-1088.

[11] 杨挺，朱希安，张帆. 基于改进三元组损失的伪造人脸视频检测方法[J]. 计算机应用研究，2021，38(12)：3771-3775.

[12] COCCOMINI D A, MESSINA N, GENNARO C, et al. Combining EfficientNet and vision transformers for video deepfake detection[C]//Image Analysis and Processing - ICIAP 2022. Cham: Springer, 2022: 219-229.

[13] TAN M X, LE Q V. EfficientNet: rethinking model scaling for convolutional neural networks[EB/OL]. [2024-01-11]. https://arxiv.org/abs/1905.11946.

[14] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. [2024-01-11]. https://arxiv.org/abs/1706.03762.

[15] WODAJO D, ATNAFU S. Deepfake video detection using convolutional vision transformer[EB/OL]. [2024-01-11]. https://arxiv.org/abs/2102.11126.

[16] RÖSSLER A, COZZOLINO D, VERDOLIVA L, et al. FaceForensics++: learning to detect manipulated facial images[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 1-11.

作者简介：
李铮，高级工程师，学士，主要从事网络与信息安全技术研究及规划工作；郑涛，工程师，硕士，主要从事网络与信息安全管理工作；张小梅，高级工程师，硕士，主要从事网络安全技术研究及管理工作。