Echoes of Authenticity : Transfer Learning and Diverse Pre-Processing Techniques

Yash Joshi,Tanush Chaudhary,Yuvraj

Electronics and Communication Engineering Deptt.
Delhi Technological University India

Abstract— **This research delves into enhancing the landscape of deepfake audio classification. Through the utilization of transfer learning techniques and a variety of pre-processing methods, significant improvements in accuracy and training efficiency are observed. The study assesses the effectiveness of various CNN-based models, including ResNet and VGG16, in conjunction with transfer learning for deepfake audio detection. Furthermore, it examines the influence of a customized data-preprocessing pipeline on improving the overall performance of the models.**

Keywords— **Transfer Learning, ResNet50, VGG16, CNN, Mel Spectrogram, Deepfake, Normalization, Trimming, Median Filtering**

I. INTRODUCTION

Deepfake technology[1], driven by sophisticated AI algorithms, has revolutionized the creation and manipulation of audiovisual content, blurring the boundaries between reality and fiction. Leveraging deep learning techniques, deepfake algorithms can generate highly convincing synthetic media, including audio recordings that mimic human speech with remarkable accuracy. However, the widespread adoption of deepfake technology has led to a surge in malicious activities, such as AI voice cloning scams, posing significant challenges to cybersecurity and trust in digital communications.

According to a survey conducted by VMware in 2022 [2], instances of deepfake attacks are escalating, with 66% of participants acknowledging their existence and increase occurrence. Among the respondents, 58% identified video as the predominant form of encountered deepfakes, while 42% cited audio. These deceptive contents were disseminated through various channels, encompassing email, mobile messaging, voice, and social media platforms.

Researchers and cybersecurity experts are continuously investigating creative approaches to identify and reduce these hazards in the face of these expanding threats. Convolutional neural networks (CNNs) are the foundation of the deepfake audio categorization landscape within this context. CNNs are very good at this kind of work because they can automatically learn data representations that are hierarchical. CNNs are able to distinguish between real and false audio recordings by capturing complex patterns and features in audio signals. This is particularly useful when it comes to audio deepfake categorization. Researchers can improve the effectiveness of deepfake detection systems by utilising transfer learning techniques in conjunction with CNN architectures. This allows researchers to take use of the rich representations that are learned from large-scale image datasets and apply them to the audio domain.

II. CONVOLUTIONAL NEURAL NETWORK (CNN)

Convolutional Neural Networks (CNNs) are a class of deep learning models that have revolutionized various fields, particularly in computer vision tasks. CNNs are designed to automatically learn hierarchical representations of data through a series of convolutional layers. These layers extract features from the input data by applying learnable filters across small regions, capturing patterns at different spatial scales. CNNs are characterized by their ability to learn spatial hierarchies of features, starting from simple features like edges and textures and progressing to more complex and abstract concepts. This hierarchical feature learning enables CNNs to effectively analyze and interpret complex visual data, making them a powerful tool for tasks such as image classification.

A. RESNET50

ResNet50[3] is a variant of the ResNet model, which stands for Residual Network and is a convolution neural network (CNN) at its depth.The number 50 came from the number of layers it compromised. It was developed by researchers at Microsoft and is widely used in the field of deep learning for tasks such as image recognition and classification. The "residual" in its name refers to the architecture's key innovation—residual blocks, which help in training much deeper networks by addressing the vanishing gradient problem through skip connections. These connections allow the network to learn an identity function, ensuring that the higher layers in the network can perform at least as well as the lower layers, thus facilitating the training of deep networks. ResNet50 is especially favored due to its balance of depth and complexity, which makes it efficient and effective for many computer vision tasks.

B. VGG16

VGG16[4] is also a deep convolutional neural network model proposed by researchers from the Visual Graphics Group at Oxford, from where its name "VGG" is derived. It features 16 layers and is recognized for its simple, yet highly effective architecture primarily consisting of convolutional layers with small receptive fields of 3x3, which are stacked on top of each other in increasing depth. Between these convolutional layers, the network utilizes max pooling to reduce spatial dimensions. VGG16 is especially recognized for its extensive use of feature maps, which grow substantially in number in the deeper layers. This enables the network to effectively capture complex patterns.

C. Transfer Learning

Transfer learning is a machine learning approach where a model designed for a specific task is repurposed to form the foundation for a different task.. It is very effective when the tasks share similarities. By utilizing the knowledge (such as features, weights, and biases) from the initial model, transfer learning enhances the learning accuracy and efficiency for the new task, especially when data availability is limited. Transfer learning is commonly used in deep learning areas such as computer vision and natural language processing. Utilizing pre-trained models from extensive datasets can substantially decrease the development time and resources needed for creating effective models for related, yet smaller-scale tasks.

ImageNet[5] is a vast dataset developed by researchers at Stanford University for training and evaluating image recognition models. It includes over 14 million labeled images that are spread across more than 20,000 distinct categories. The organization of the dataset follows the WordNet hierarchy, with each category in the hierarchy represented by hundreds or even thousands of images. ImageNet has significantly contributed to improvements in Machine learning and Deep learning algorithms that can achieve high accuracy in visual recognition tasks and is a cornerstone dataset in AI research.

III. PROBLEM STATEMENT

Recurrent neural networks (RNNs), convolutional neural networks (CNNs), or a combination of the two, are frequently employed in DL-based techniques. CNNs have shown promise in audio classification tasks, especially when combined with visual audio representations like spectrograms and histograms. One such CNN that was specifically designed for picture identification applications is the residual neural network (ResNet). Furthermore, transfer learning has been shown to improve performance on some tasks such as deepfake detection and audio categorization.

Building upon these insights, this paper will concentrate on employing transfer learning for audio deepfake detection using a ResNet architecture as a core component. The investigation will delve into four subcategories:

Model Performance: Evaluating the performance of the models.

Impact of Transfer Learning: Examining if transfer learning improves performance compared to randomly initialized weights.

Model Comparison: Comparing models using ResNet50 and VGG16 architectures.

Preprocessing Evaluation: Assessing the effectiveness of preprocessing techniques on model performance.

IV. DATASET UTILISED

In-the-Wild[6] dataset introduced in 2022 by Müller et al. is specifically designed to evaluate the performance of audio deepfake detection models in real-world circumstances. Distinguished from other datasets by its emphasis on public figures, it encompasses audio content sourced from various online platforms, reflecting authentic encounters one might have in daily life.The dataset includes recordings of 58 celebrities and politicians, featuring notable figures such as Arnold Schwarzenegger, Queen Elizabeth II, and Barack Obama, among others.

In total the dataset has 17.2 hours of deepfake audio mixed in with 20.8 hours of real audio. Every speaker contributes roughly 23 minutes of real audio and 18 minutes of deepfake audio on average. While the dataset maintains standardization in terms of sample rate and file format, minimal preprocessing has been applied.

We chose to use the In-the-Wild dataset for our research since it offers a decent amount of English-language labeled audio data. Furthermore, the minimal preprocessing of the dataset gave us the flexibility of experimenting with various kinds of preprocessing methods. Additionally, the size of the dataset was manageable, which helped in mitigating computation complexity when compared to other larger datasets.

V. DATA ANALYSIS AND VISUALISATION

The data comprises 31,779 audio files of varying lengths, categorized into 11,816 spoof samples and 19,963 bona-fide samples. To facilitate model training, testing, and validation, we adopted a data splitting strategy where the dataset was divided into three subsets: training, testing, and validation, in the ratio of 65:20:15, respectively.

Additionally, as the models under investigation rely on image inputs, we converted the audio files into corresponding spectrograms. This transformation allowed us to visually represent the audio data, making it compatible with Convolutional Neural Network (CNN) architectures. Notably, CNNs serve as the foundational framework for all models under study, including ResNet, VGG16, and Xception.

VI. PREPROCESSING TECHNIQUES

We aimed to investigate the impact of various kinds of preprocessing on the model's performance by contrasting the results obtained when training on raw spectrograms with those achieved after processing the data through a comprehensive preprocessing pipeline. This pipeline, outlined in detail in the following paragraphs, encompasses various techniques designed to enhance the quality and relevance of the input data for machine learning tasks.

Normalization: It is a preprocessing technique aimed at standardizing the amplitude of audio signals within a fixed range to facilitate consistent processing by machine learning models. By scaling the amplitude to have unit norm [-1:1], we ensure uniformity across different audio samples, thereby mitigating issues related to signal intensity variations. The dynamic range of audio signals is effectively adjusted, leading to improved model convergence and generalization. This technique served as the fundamental step in our preprocessing pipeline, laying the groundwork for subsequent processing stages.

Trimming:This step helped us in eliminating irrelevant silence or low-energy segments from audio signals. By detecting and discarding segments with energy levels below a specified threshold (30 dB from the RMS value of the audio in our case), it enhanced the relevance of the input data by focusing on meaningful signal segments. This technique also helped reduce computational overhead by excluding non-informative segments.

Median Filtering: This is a noise reduction technique widely employed in audio preprocessing to suppress short-duration noise bursts or impulsive artifacts. By replacing each data point with the median value within a specified window, median filtering effectively helped us smooth out sudden variations or spikes in the audio signal, facilitating robust feature extraction and classification. Moreover, median filtering helps mitigate the influence of outliers or anomalies, resulting in a more stable and consistent representation of the signal.

Mel Spectrograms: It is a representation of audio signals which is computed by applying a Mel filterbank to the magnitude spectrum of a signal, which results in a spectrogram where the frequency axis is scaled according to the Mel scale.

It is a perceptual scale of pitches that is based on the human auditory system's response to different frequencies which designed to reflect the way humans perceive pitch differences, with higher resolution at lower frequencies and lower resolution at higher frequencies.It allowed us to capture important acoustic features of audio signals, such as pitch and timbre, in a compact and efficient representation.

No . of epoch	Accuracy	Loss	Time To Train (seconds)
1	0.9330	0.2131	339.45
2	0.8430	0.3697	656.87
5	0.9524	0.1640	1759.90

Table 1 : ResNet50 Performance with Transfer Learning but without any pre-processing applied.

No . of epoch	Accuracy	Loss	Time To Train (seconds)
1	0.9825	0.0698	915.18
2	0.9057	0.2611	1654.16
5	0.9825	0.0528	3836.31

Table 1 : ResNet50 Performance with Transfer Learning and pre-processing.

No . of epoch	Accuracy	Loss	Time To Train (seconds)
1	0.7465	0.1631	774.45
2	0.7515	0.4520	1510.60
5	0.8686	0.1046	3606.57

Table 3 : VGG16 Performance with Transfer Learning and no pre-processing

VII. RESULTS

Considering the progress made with transfer learning and pre-processing techniques, the project has shown promising results in improving accuracy and reducing training time. It's clear that the combination of residual neural networks and transfer learning is effective for detecting deepfake audio, although there are certainly other potential methods worth exploring.

VIII. CONCLUSION

This project delved into the realm of audio deepfake detection utilizing machine learning techniques. Through an extensive review of existing literature, it became evident that both residual neural networks and transfer learning are effective methodologies in this domain. Subsequently, a novel approach for audio deepfake detection was proposed. This approach involves converting audio signals into spectrograms and inputting them into a comprehensive preprocessing pipeline.

The ResNet50 and VGG16 model, originally pre-trained on a comprehensive dataset of natural images, was further fine-tuned for audio classification using a specialized dataset.

ResNet50 performed better in terms of accuracy than VGG16, according to a comparative analysis, while transfer learning improved total accuracy while cutting training time. A proper preprocessing pipeline's implementation also showed a notable improvement in the performance of the model.

IX. FUTURE SCOPE

In future research, using bigger datasets like WaveFake with more languages and longer content can help improve the results. Also, by using more advanced models with extra layers, Using more advanced models with extra layers, like the RESNET150 model, can provide a more detailed learning process. This can lead to enhanced results in our research. To make things even better, researchers can train the models with higher training steps and epoch, which will make the results more accurate and reliable. These changes can lead to better applications in different areas.

X. APPLICATION

This technology can have a wide variety of applications in the current market some of which are:

Strengthens security measures by identifying AI-generated audio in authentication systems, preventing unauthorized access to sensitive data.

Aids law enforcement and forensic experts in the investigation of cybercrimes involving synthetic audio evidence.

Enable fact-checking organizations to verify information and counteract false narratives and combat the spread of propaganda.

References

[1] Z. Khanjani, G. Watson, and V. P. Janeja, “Audio deepfakes: A survey,” 2023. DOI: 10.3389/fdata.2022.1001063

[2] VMware, “VMware Global Incident Response Threat Report 2022,” 2022.

[3] He, Kaiming & Zhang, Xiangyu & Ren, Shaoqing & Sun, Jian. (2015). Deep Residual Learning for Image Recognition. 7.

[4] Simonyan, Karen & Zisserman, Andrew. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv 1409.1556.

[5] J. Deng, W. Dong, R. Socher, L. -J. Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, 2009, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.

[6] Müller, Nicolas & Czempin, Pavel & Dieckmann, Franziska & Froghyar, Adam & Böttinger, Konstantin. (2022). Does Audio Deepfake Detection Generalize?.

[7] F. Tom, M. Jain, and P. Dey, “End-To-End Audio Replay Attack Detection Using Deep Convolutional Networks with Attention,” Interspeech 2018, 2018. DOI: 10.21437/Interspeech.2018-2279.

[8] *Keras Team, “Keras Applications,” [Online]. Available: https://keras.io/ api/applic ations/, Last accessed: 29/05/2023*