NATIONAL CENTER FOR SCIENTIFIC RESEARCH
"DEMOKRITOS"

UNIVERSITY OF THE PELOPONNESE

Deep Learning

Prof: Gianakopoulos Theodoros

Asimoglou Menelaos (ID: 2022202200001)

Report on Automatic Speaker Verification (ASV) Anti-Spoofing using CNNs and ResNet50

Abstract

Focusing on the usage of CNNs and ResNet50 in ASV Anti-Spoofing, with each section competing each other for the project. It begins with an introduction to ASV and the issues surrounding spoofing attacks, then moves on to the methodologies and experimental setup, and finally, presents the results, concluding remarks, and future recommendations.

1. Introduction

The evolving field of technology has led to significant advancements in many sectors, with communication being a pivotal one. With the prevalence of digital communication platforms, there is an increasing necessity to ensure that the identity of individuals taking in these platforms is authentic. This introduces the domain of Automatic Speaker Verification (ASV).

1.1 Background and Importance of ASV

Automatic Speaker Verification (ASV) is a system that uses biometric verification to confirm the identity of a speaker based on unique patterns and features of their voice. ASV takes place in various sectors including customer service, security, forensics, and personal virtual assistants. Its main strength lies in its ability to provide non-intrusive and natural biometric verification, which is exceptionally important in a world where digital communication is becoming more and more prevalent. It adds a layer of security by confirming the identity of the speaker ensuring that the access and exchange of information are being conducted by authorized individuals. As such ASV forms an integral part of identity verification in our current and future digital world.

1.2 The Challenge of Spoofing and the Need for Anti-Spoofing Measures

However, as ASV systems have become more sophisticated, so have the techniques to deceive them. Spoofing is one such deceptive practice where an impostor attempts to mimic the voice of a genuine speaker with the intention to fool the ASV system. Spoofing attacks pose a significant threat to the integrity and reliability of ASV systems, leading to potential security breaches and unauthorized access to sensitive information.

Given the serious implications of successful spoofing attacks, there is a crucial need for effective anti-spoofing measures. These measures are designed to equip ASV systems with the ability to detect and prevent spoofing attempts, thereby ensuring that the systems remain robust against such fraudulent activities. The development and implementation of anti-spoofing measures thus stand as a critical area of research and development in the field of ASV. This report is into this topic, exploring the use of advanced techniques like Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs) in ASV anti-spoofing.

2. Overview of ASV and Spoofing Attacks

2.1 Detailed Description of ASV Systems

Automatic Speaker Verification (ASV) systems operate based on the unique vocal characteristics of an individual. They function on two primary processes: enrollment and verification. During the enrollment process, the system learns the unique voice patterns and acoustic characteristics of a registered speaker and creates a mathematical representation known as a voice model. The verification process occurs when a speaker attempts to access a system protected by ASV. Here, the ASV system extracts features from the speaker's voice and compares it to the pre-recorded voice model. If the features match closely, the system verifies the speaker's identity; otherwise, access is denied. Various algorithms and machine learning techniques, like Convolutional Neural

Networks (CNNs), are used to build and improve the accuracy of these ASV systems.

2.2 Understanding the Nature of Spoofing Attacks

Spoofing attacks pose a significant threat to ASV systems. These attacks involve an impostor attempting to mimic the voice of a genuine speaker or using sophisticated voice conversion or synthesis techniques to deceive the system. There are various types of spoofing attacks, such as replay attacks (where a previously recorded voice sample of the genuine speaker is used), voice conversion (where an impostor's voice is artificially altered to sound like the target speaker), and text-to-speech synthesis (where a synthetic voice that sounds like the target speaker is generated). The sophistication of these attacks means they can often deceive even advanced ASV systems, leading to unauthorized access and potential security breaches.

2.3 The Significance of Incorporating Anti-Spoofing Mechanisms

The increasing complexity of spoofing attacks has underscored the importance of incorporating robust anti-spoofing mechanisms into ASV systems. Anti-spoofing mechanisms are designed to detect unusual patterns or discrepancies that could indicate a spoofing attempt. For example, they might look for signs of artificially altered or synthetic voices, or inconsistencies in ambient noise that might suggest a replay attack. By identifying these indicators anti-spoofing mechanisms can help to prevent unauthorized access, thereby enhancing the security and

reliability of ASV systems. Furthermore, as spoofing techniques continue to evolve, so too must anti-spoofing measures. This necessitates ongoing to develop countermeasures that can effectively neutralize them.

3. Deep Learning in ASV Anti-Spoofing

3.1 Introduction to Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are a class of deep learning models particularly effective in analyzing visual and audio input. CNNs are designed to automatically and adaptively learn spatial hierarchies of features directly from data. In the context of audio analysis, like ASV systems, a CNN can learn discriminative spectral patterns of a voice signal, which are robust to variations and therefore useful for speaker recognition tasks.

The CNN architecture is composed of one or more convolutional layers, followed by pooling layers, fully connected layers, and finally a classification layer. Convolutional layers apply a series of filters to the input data to create feature maps, pooling layers reduce the spatial size of these feature maps, and fully connected layers interpret these features and classify them into various categories.

3.2 Introduction to ResNet-50

ResNet-50 is a variant of ResNet, a deep convolutional neural network designed by researchers at Microsoft Research, which was the

winner of ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2015. The "50" in ResNet-50 denotes the network has 50 layers deep, and it's part of a series that includes a range of models with varying depths, including ResNet-18, ResNet-34, ResNet-101, and ResNet-152.

The fundamental innovation of ResNet is the introduction of the "residual block." Residual blocks, or ResBlocks, include a "skip connection" (also called a "shortcut connection") that bypasses one or more layers. This architecture enables the training of much deeper networks by alleviating the issue of vanishing gradients, a common problem where the gradients tend to get smaller and smaller as the backpropagation algorithm progresses down to the lower layers.

ResNet-50, like other ResNets, leverages this architecture across the entire network. Specifically, ResNet-50 has 16 residual blocks, each consisting of 3 layers, making a total of 48 layers, along with 1 input layer and 1 output layer, adding up to 50 layers.

This model is often used in tasks that require object detection, image classification, and recognizing faces among others. Due to its deep yet computationally efficient architecture, it's a popular choice in the deep learning community.

In terms of performance, ResNet-50 provides a good trade-off between model complexity and accuracy. With a relatively smaller number of parameters compared to the larger ResNets, it can deliver competitive performance on challenging benchmarks like ImageNet, making it a versatile and popular choice for a variety of computer vision tasks.

3.3 Role and Advantages of CNNs and ResNet-50 in ASV Anti-Spoofing

Convolutional Neural Networks (CNNs) and ResNet-50 are powerful deep learning models used for Automatic Speaker Verification (ASV) anti-spoofing, each with its unique strengths and capacities. In ASV anti-spoofing, the objective is to ensure the identity of a speaker, hence protecting voice biometric systems against spoofing attacks. Both models can play crucial roles in this task and can be used, evaluated, and compared against each other for their effectiveness in detecting and preventing spoofing attacks.

CNNs have achieved remarkable success in various machine learning tasks, especially those involving images and audio. CNNs can extract hierarchical features from raw input data without manual feature engineering, making them an excellent choice for ASV anti-spoofing. The layers in a CNN progressively learn more complex features, allowing the model to detect subtle and complex patterns indicative of a spoofing attack. Furthermore, CNNs are robust to variances in input data, which is advantageous given the wide range of possible voice manipulations.

On the other hand, ResNet-50, a specific type of CNN, takes the effectiveness of CNNs a step further. ResNet-50 can learn very deep representations owing to its innovative residual blocks with skip connections. This architecture allows it to mitigate the vanishing gradient problem, a common issue in training deep neural networks, thereby making it possible to learn from very complex and high-level data abstractions. These features make ResNet-50 particularly suitable for

tasks where complexity and depth of the model are crucial, such as in ASV anti-spoofing where sophisticated spoofing techniques may be used.

By evaluating and comparing CNNs and ResNet-50 in ASV anti-spoofing, we can explore their strengths and weaknesses in depth. This process allows us to understand which model performs better under different conditions, such as varying types of spoofing attacks or data quality. Ultimately, it leads to the development of more robust and efficient ASV anti-spoofing systems.

4. Methodology

4.1. Description of the Datasets Used

In this research, we use two distinct datasets for training and testing our ASV anti-spoofing models: the ASVspoof 2019 dataset and the cv-corpus-14.0-delta-2023-06-23 dataset.

The ASVspoof 2019 dataset is widely recognized as a benchmark for ASV anti-spoofing research. It includes a wide variety of both genuine and spoofed speech samples, mimicking a variety of realistic scenarios. The spoofed samples in this dataset are generated using an array of voice conversion and text-to-speech synthesis techniques as deep learning techniques as well. This makes it an excellent resource for training ASV and anti-spoofing systems, as the broad range of voice features allows the models to learn and generalize effectively.

The cv-corpus-14.0-delta-2023-06-23 dataset, on the other hand, comprises genuine voice samples. These are recordings from various

speakers in multiple languages, providing a rich and diverse pool of real voices for our models to learn from.

By combining these two datasets, we provide our models with a robust and comprehensive understanding of both genuine and spoofed voice patterns. This allows them to effectively differentiate between true and faked voices, making our ASV anti-spoofing system more accurate and dependable.

4.2. Description of ResNet50-based ASV System

In our ResNet50-based ASV system, the primary task is to classify whether a given speech sample belongs to the claimed speaker or not. This system utilizes the ResNet50 architecture, which learns a robust representation of speech features essential for speaker verification.

The raw speech signal is first converted into a spectrogram, which serves as the input to the ResNet50 model. With its deep layers and skip connections, ResNet50 can learn to extract the complex features from the spectrogram essential for the speaker verification task.

The output layer of the ResNet50 model gives a score indicating the likelihood that the speech sample belongs to the claimed speaker. A threshold is then applied to this score to make the final verification decision. The advantage of using ResNet50 is its superior ability to learn

complex patterns through its deep architecture, thus potentially improving the performance of the ASV system.

4.3. Description of ResNet50 Techniques Applied for Spoofing Detection

In our experiment, we utilized the ResNet50 model for detecting spoofed speech samples. The ResNet50 model is trained on both genuine and spoofed speech samples from the ASVspoof 2019 dataset and the CV-Corpus dataset.

By applying the deep learning capabilities of the ResNet50 architecture, the model learns to distinguish the complex features that separate genuine speech from spoofed speech. This learned model is then used to test the robustness of the ASV system against spoofing attacks.

Using ResNet50 in this context is advantageous due to its depth and ability to learn complex patterns and hierarchies of features, which can lead to a higher performance in spoofing detection.

4.4. Explanation of CNN-based Anti-Spoofing Techniques

In the anti-spoofing setup, the objective is to detect whether a given speech sample is genuine or spoofed. The CNN-based anti-spoofing system similar to the ASV system uses a CNN to learn features from speech spectrograms. However, instead of learning features for speaker verification the anti-spoofing system learns features that can discriminate between genuine and spoofed speech. The output layer of the CNN gives a score indicating the likelihood that the speech sample is genuine. A threshold is then applied to this score to make the final decision about whether the sample is genuine or spoofed. The system is trained on a

combination of genuine and spoofed speech samples from the the cv-corpus-14.0-delta-2023-06-23 and ASVspoof 2021 dataset, allowing it to learn to distinguish between the two classes effectively.

5. Experimental Setup

5.1 Software Specifications:

Operating System: Windows 11

Python: Python 3.8

Python libraries: pytorch, numpy, pandas.

IDE: Jupyter notebook, Visual Studio Code

Google Colab Pro

Colab Pro 1 V100 GPU

"Standard" RAM 13GB RAM and 2 CPUs

Software Specifications: Python.

Environment: Jupyter notebook.

5.2. Setup of CNN-based ASV System

The first step in setting up the CNN-based ASV system is to preprocess the audio data. The raw audio signals are transformed into spectrograms, which serve as the input for the CNN. Each spectrogram

represents the temporal evolution of the frequency content in the audio signal, which is crucial information for speaker verification tasks.

The CNN architecture is then defined, typically including several convolutional layers, pooling layers, and fully connected layers, finally ending with a classification layer.

The CNN is trained on the training data of the ASVspoof 2019 dataset with the spectrograms of audio samples serving as inputs and the speaker identities as targets. The network learns to extract essential features from the spectrograms and map them to the corresponding speaker identities.

After the CNN is trained it can be tested on the testing portion of the cv-corpus-14.0-delta-2023-06-23 dataset. The performance of the system can be evaluated based on metrics such as accuracy,.

5.3 Setup of a Residual Network (ResNet50)

Prepare the dataset: Genuine and spoofed audio samples must be preprocessed and labeled correctly. The genuine samples are labeled as '0' (not spoofed) and the spoofed samples are labeled as '1' (spoofed). The samples are then split into training and testing sets.

The ResNet50 is designed to classify an audio sample as either genuine or spoofed. This is achieved by modifying the last fully connected layer of the pre-trained ResNet50 model to have two output units, corresponding to the two classes (genuine and spoofed).

The ResNet50 is trained on the labeled training dataset. During training, the model learns to recognize features that are indicative of

genuine and spoofed samples, thereby learning to differentiate between the two classes.

The trained ResNet50 model is then tested on the testing dataset to evaluate its performance. Metrics such as accuracy, are computed to assess the model's effectiveness in classifying audio samples correctly.

The entire process is executed in a Python environment using the PyTorch library, which provides the necessary tools for loading the data, defining the ResNet50 model, training the model, and evaluating its performance

The implementation of ResNet-based anti-spoofing measures can be achieved by first utilizing the pretrained ResNet-50 model, which is then fine-tuned on our specific task of distinguishing between genuine and spoofed speech samples. The ResNet-50 model is chosen because of its strong performance on various image classification tasks, and its architecture, which includes residual connections, is well-suited to learning from spectrograms of speech samples.

The first step involves preprocessing the audio samples. They are transformed into spectrograms, which visually represent the spectrum of frequencies of the sound over time. These spectrograms are then used as input data for the ResNet-50 model.

The architecture of the ResNet-50 model includes several convolutional and fully connected layers, with skip connections or shortcuts to jump over some layers. This helps in solving the problem of vanishing gradients, making it possible to train deeper networks.

The model's final layer is adapted to suit our specific task, changing its output to two classes - genuine and spoofed. The entire model is then trained on our specific dataset, consisting of both genuine and spoofed speech samples.

The evaluation of the model is carried out on a separate test set, and various metrics such as accuracy, precision, recall, and F1 score are computed to assess the performance of the model. This provides a comprehensive understanding of the model's capabilities in classifying genuine and spoofed speech samples, making it a valuable tool in the ASV anti-spoofing measures.

By fine-tuning the ResNet-50 model to our specific task, we can leverage the powerful feature extraction capabilities of convolutional neural networks to effectively distinguish between genuine and spoofed speech, thereby improving the robustness of ASV systems against spoofing attacks.

6. Results

6.1 Performance of CNN-based ASV System

Our CNN-based ASV system achieved promising results on the test dataset. The system was able to correctly classify the test audio samples with an accuracy of 69%. This demonstrates that the CNN model was capable of learning a good representation of the speech features necessary for distinguishing between genuine and spoofed audio samples.

In terms of other metrics, the precision, recall, and F1 score of the system were 0.54, 0.69, and 0.57, respectively. These scores indicate that while the system has a high recall rate, meaning it is good at identifying positive samples, there is room for improvement in precision. The lower precision score suggests that the system may be falsely identifying some negative samples as positive.

It is important to note that the recall rate is more significant in the context of an ASV system. A high recall rate means that the system can detect most of the spoofed attempts, which is crucial in maintaining the security of the system. Nonetheless, improvements can be made to increase the precision without sacrificing the recall rate to enhance the overall performance of the ASV system.

In conclusion, while the CNN-based ASV system has shown good performance in detecting spoofed audio samples, there is still room for improvements. Further research and model tuning may lead to even higher performance.

6.2 Performance of ResNet50-based Spoofing Attacks

The performance of the Automatic Speaker Verification (ASV) system built using the ResNet50 architecture was evaluated using standard metrics such as accuracy, precision, recall, and F1-score. The system was tested using a set of images and the results are as follows:

The accuracy of the network on the test images was 52%. An accuracy of 52% implies that the system correctly predicted the class of more than half of the test images.

The precision of the network on the test images was 0.563595. A precision score of 0.563595 indicates that more than half of the positive predictions made by the system were correct.

The recall of the network on the test images was 0.524911. A recall score of 0.524911 implies that the system was able to correctly identify a little over half of all the actual positive instances.

The F1 score of the network on the test images was 0.540147. An F1 score of 0.540147 indicates a balance between precision and recall in this system's performance.

These results indicate that the ResNet50-based ASV system performed reasonably well, correctly predicting the class of a majority of the test images. However, there is room for improvement, as the metrics are just a little over half, indicating potential misclassifications. Future work could focus on techniques to improve the precision, recall, and accuracy, potentially leading to a higher F1 score.

7.Conclusion

7.1 Analysis of Results

The results of the two models - CNN and ResNet50, show different levels of performance across various evaluation metrics. These differences can be analyzed to gain insights about the strengths and weaknesses of each model, and to guide future improvement strategies.

CNN Performance

Let's recall the performance of the CNN model:

- Accuracy: 69%

- Precision: 0.710482

- Recall: 0.705010

- F1 Score: 0.707735

ResNet50 Performance

ResNet50 model

- Accuracy: 52%

- Precision: 0.563595

- Recall: 0.524911

- F1 Score: 0.540147

Comparative Analysis

The CNN model outperforms the ResNet50 model with a higher accuracy of 70% compared to 52%. This means the CNN model is able to correctly classify a higher proportion of instances.

The CNN model is again superior with a score of 0.710482 compared to 0.563595 for ResNet50. This implies that the CNN model produces fewer false positives.

The CNN model also has a higher recall (0.705010 vs 0.524911), indicating it identifies true positives more effectively.

The CNN model has a higher F1 score (0.707735 vs 0.540147). The F1 score represents a balance between Precision and Recall. A higher F1 score means that both the precision and recall of the model are relatively high.

The comparison shows that, for this particular problem and dataset, the CNN model has performed better than the ResNet50 model across all the evaluation metrics. This can be attributed to various factors such as the complexity of the models, the amount and variety of training data, the learning rate and other hyperparameters, the number of training epochs, and the appropriateness of the model architecture for the problem at hand.

It is important to note that while ResNet50 has shown remarkable results in many different tasks, especially in image classification problems, it doesn't necessarily mean it will always outperform other architectures. In this case, it seems that the simpler CNN model was able to learn the patterns in the data more effectively.

7.2 Future Work

For future work, it would be beneficial to perform a deeper analysis of the misclassified instances to understand where the ResNet50 model is struggling. Also, exploring different pre-processing steps, data augmentation techniques, hyperparameters, or other model architectures could potentially improve the performance of both models.