Technical Review: Siamese Convolutional Neural Network for Audio Deepfake Detection

Kartik Dua

05/04/2025

Abstract

This document presents a comprehensive technical review of the Siamese Convolutional Neural Network (SCNN) designed and implemented for detecting audio deepfakes using the DECRO dataset. The model architecture incorporates residual connections and squeeze-and-excitation blocks to enable robust feature learning from spectrogram representations of audio data. We elaborate on the theoretical foundation, code implementation decisions, and insights gathered during experimentation.

1 Introduction

The prevalence of deepfake audio calls for reliable speaker verification systems that can distinguish between authentic and manipulated voices. Siamese neural networks have demonstrated effectiveness in verification tasks by learning similarity metrics in an embedding space. Inspired by the original work of Bromley et al. (1993) on signature verification [1], we designed a refined Siamese CNN architecture tailored to audio spectrogram data.

Our approach is also motivated by advancements in computer vision and audio classification, leveraging architectures such as ResNet [?] and Squeeze-and-Excitation Networks [?].

2 Theoretical Foundation

2.1 Siamese Neural Networks

A Siamese network consists of two identical subnetworks that share weights and learn to project input data into a latent space. The network optimizes a contrastive loss function that minimizes the distance between embeddings of similar inputs and maximizes it for dissimilar pairs.

2.2 Residual Connections

Residual connections help mitigate the vanishing gradient problem in deep networks by providing shortcut paths for gradients to flow backward. This facilitates training of deeper models without significant degradation in performance.

2.3 Squeeze-and-Excitation (SE) Block

SE blocks adaptively recalibrate feature maps by explicitly modeling interdependencies between channels. This enhances representational power and helps the model focus on informative features.

2.4 Contrastive Loss

Given a pair of embeddings (x_1, x_2) and a binary label y, the contrastive loss is defined as:

$$\mathcal{L}(x_1, x_2, y) = (1 - y) \cdot \|x_1 - x_2\|^2 + y \cdot \max(0, m - \|x_1 - x_2\|)^2 \tag{1}$$

where m is the margin that defines the separation boundary for dissimilar pairs.

3 Implementation Breakdown

3.1 Architecture Design

The model begins with an initial convolutional layer followed by three residual blocks of increasing depth. Each residual block contains two convolutional layers with batch normalization and ReLU activations, followed by a shortcut connection.

- **Initial Convolution:** Extracts low-level features from the input spectrogram.
- **Residual Blocks:** Three blocks with channel sizes 32, 64, 128, and 256.
- **SE Block:** Applied after the final residual layer to perform channel-wise attention.
- **Fully Connected Layers:** Final embeddings are 256-dimensional and learned via a two-layer MLP.

3.2 Code Explanation

- `ResidualBlock`: Contains two convolutions and a shortcut connection. Used to stabilize training and encourage reuse of features.
- `SEBlock`: Implements global average pooling followed by two fully connected layers and a sigmoid activation for recalibration.
- `RefinedSiameseCNN`: Combines all modules into a unified architecture. During the first forward pass, it initializes the FC layer dimensions dynamically.

- `ContrastiveLoss`: Calculates pairwise distances and applies the margin-based penalty.
- `train()` and `validate()`: Functions handle model training and threshold sweeping for optimal decision boundary.

3.3 Training Configuration

- Batch Size: 32
- Epochs: 40
- Optimizer: Adam (learning rate 0.0005)
- Margin: 2.0
- Input: 1-channel spectrogram (128×128)

4 Insights and Observations

- Validation accuracy varied significantly based on the distance threshold. A sweep from 0.2 to 2.1 (step 0.05) was performed.
- The SE block improved performance by helping the model focus on speaker-relevant channels.
- Dynamic FC layer initialization ensured compatibility with input dimensions without hardcoding.

5 Conclusion

This SCNN implementation demonstrates the effective use of contrastive learning for audio verification tasks. The combination of residual learning and SE attention mechanisms makes the model both deep and context-aware. The methodology provides a robust baseline for future enhancements like transformer-based attention or multimodal fusion.

References

[1] Bromley, J., Guyon, I., LeCun, Y., Sackinger, E., & Shah, R. (1993). Signature verification using a "Siamese" time delay neural network. In *Neural Information Processing Systems (NIPS).*

[2] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[3] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*