Spectrogram-Based Gunshot Detection using CNNs and Dual-Headed EffNet

Ashwin Kumar Udayakumar, Karthikeyan Shanmugam, Stanley Jovel

{au2177, ks6964, lsj3272}@nyu.edu

Abstract

In this paper, we present a vision-based approach to gunshot sound recognition designed to classify and detect gunshots amidst diverse background noises in urban environments such as parks, malls, and airports. Our methodology involves the insertion of gunshot sounds into background noise samples at random intervals and with varying volumes. These audio samples are then converted into spectrograms to serve as input data for our models. We explore and evaluate the performance of four distinct neural network models: a custom Convolutional Neural Network (CNN), a ResNet18, a custom Dual-Headed EfficientNetV2 (EffNetV2), and a Vision Transformer (ViT), each trained and validated on the above dataset. Our results demonstrate that these models, especially the EffNetV2 and the ViT, show promising capabilities in accurately detecting gunshots, with 85-88% accuracy achieved under various noise conditions. Code for this paper is available at: https://github.com/Stanley-Jovel/Computer-Vision-Project

I. Introduction

The application of computer vision techniques extends beyond traditional image analysis, as demonstrated by their effectiveness in extracting meaningful information from various types of visual inputs. Studies like "Bird Sound Recognition Using a Convolutional Neural Network"[1] highlight computer vision methods' versatility in interpreting not only images but also sound.

Audio signals are inherently one-dimensional, and to leverage vision-based models for audio processing, a representative transformation of the audio to a two-dimensional form is required. Spectrograms, which visually depict the spectrum of frequencies in a sound signal as they vary with time, serve as an ideal two-dimensional representation. These are generated using the Short Time Fourier Transform (STFT)[2], effectively translating audio into a format amenable to image-based analysis.

The focus of this paper is on the development of an effective model for the recognition and classification of gunshot sounds, a critical task in public safety and security monitoring. However, a significant challenge encountered was the absence of a reliable, comprehensive dataset of gunshot audio, which is essential for training and validating such models. To address this, we created a synthetic dataset by programmatically embedding gunshot sounds into various background noise samples, simulating urban environments like parks, malls, and airports. This approach, while innovative, presents several limitations:

1. Limited Real-World Variability: A synthetic dataset may not capture the full spectrum of real-world acoustic variations and nuances, potentially limiting the model's adaptability to diverse environments.
2. Generalization Challenges: Models trained on synthetic data might struggle to generalize well to unseen, real-world scenarios.
3. Potential Bias: The process of generating the dataset may introduce unintentional bias, influenced by the selection and characteristics of the gunshot sounds and background noises.

These considerations are vital for understanding the scope and potential limitations of our study. Readers should keep these factors in mind when interpreting the results and conclusions drawn from our experiments.

II. Methodology

A. Dataset creation

Our dataset is a collection of over 11,000 spectrogram images in PNG format, derived from 10-second audio clips. These clips incorporate a mix of gunshot sounds, similar-sounding non-gunshot noises, and pure background audio to create a diverse training environment for our models. The dataset generation process is as follows:

1. **Background sounds**: We collected ambient audio recordings from diverse environments to simulate real-world scenarios. Sources included YouTube and SoundCloud, capturing everyday sounds such as traffic, human chatter, nature sounds, and urban activity. Examples include *alberta_mall.mp3, central_park.mp3, and tokyo.mp3*.
2. **Audio segmentation**: We segmented each background recording into discrete 10-second slices to standardize the audio input length.
3. **Gunshot selection**: For each segment, we introduced a stochastic element where there is an 80% chance of selecting a gunshot sound and a 20% chance of selecting a non-gunshot sound for later insertion. Non-gunshot sounds mimic gunshots to a degree and include *car_backfire.ogg, balloon_bursting.wav, and fireworks.wav.*
4. **Volume variations**: We produced three distinct volume levels – 8%, 35%, and 70% – for each chosen gunshot sound to account for distance and environmental sound dampening effects.
5. **Gunshot insertion**: The gunshot sounds, with their volume variations, were then inserted into the audio segments at random start times. This resulted in three unique renditions of each background segment, each containing the same gunshot sound at different volumes and starting points.
6. **Spectrogram generation**: For each audio variation, a corresponding spectrogram was generated. An additional spectrogram for each segment without any insertions was also produced.

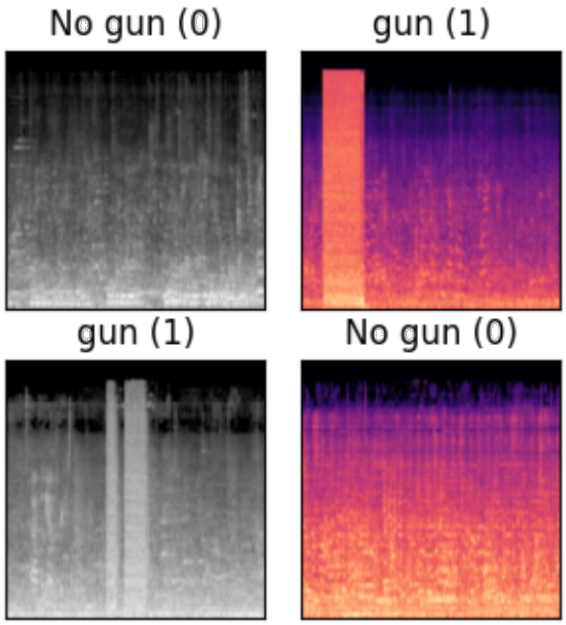

Figure 1. Example of the Spectrograms in our Dataset: On the left, two samples are shown in grayscale, depicting the audio without a gunshot ("No gun (0)") and with a gunshot ("gun (1)"). On the right, the viridis color scheme is used to represent similar audio events.

The dataset was generated in two different color schemes: viridis and grayscale (see Figure 1). They were not mixed and were used separately in experiments (further elaborated in the Experiments section).

B. Models

This study evaluates four distinct neural network architectures, each chosen for their potential efficacy in spectrogram-based gunshot detection. The comparative analysis of these models aims to identify the most effective approach for this task.

B.1 Custom Model

We developed a custom convolutional neural network (CNN) featuring two convolutional layers. Considering the relatively modest size of our dataset (over 11,000 samples), we opted for a shallow network architecture to mitigate the potential for overfitting. The model incorporates Rectified Linear Units (ReLU) to introduce non-linearity,

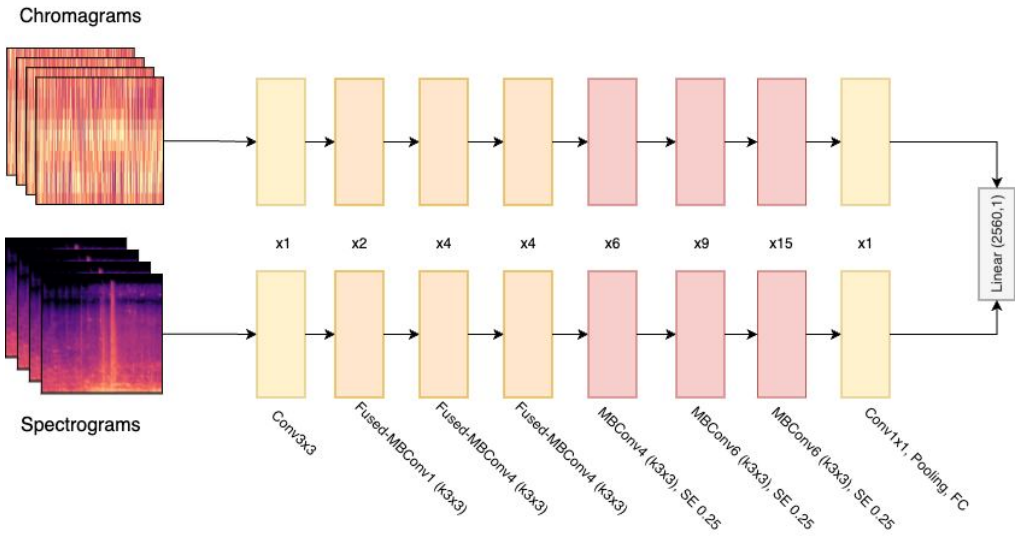

Figure 2. Architecture of the two headed EffNetV2 network implemented. One head is trained using spectrograms and the other is trained using Chromagrams.

MaxPooling for dimensionality reduction, Batch Normalization to aid in stabilizing the training process, and a dropout layer to further prevent overfitting. The output is a single neuron for binary classification, determining the presence or absence of a gunshot in the input spectrogram. The model architecture is as follows:

Input

↳ Conv2d(3,64,3x3,padding=1)

ReLU

MaxPool2d(2x2)

BatchNorm2d(64)

↳ Conv2d(64,128,3x3,padding=1)

ReLU

MaxPool2d(2x2)

BatchNorm2d(128)

↳ Flatten

↳ Linear(56*56*128, 512)

ReLU

Dropout

↳ Linear(512, 1)

Output

B.2 ResNet18

For our experiments, we utilized the ResNet18 model readily available within the PyTorch library. This choice was motivated by our hypothesis that shallower architectures may yield better results for our dataset's size and complexity. To validate this, we aimed to compare the performance of ResNet18, a relatively deeper model, against our custom shallow CNN.

Furthermore, we investigated the model's performance both with and without pre-trained weights. This comparison was to ascertain whether the ImageNet pre-trained weights, given their dissimilarity to our dataset of spectrograms, hinder or enhance the model's accuracy.

B.3 Dual-Headed EffNetV2

In our experiments, we identified that both Chromagrams and Mel spectrograms individually provide valuable insights into the presence of gunshots. To leverage the strengths of both, we customized the EfficientNetV2[3] (EffNetV2) architecture (a state-of-the-art model) to include dual parallel feature layers – one dedicated to Mel spectrograms and the other to Chromagrams.

This dual-headed approach allows the model to simultaneously process and learn from these two

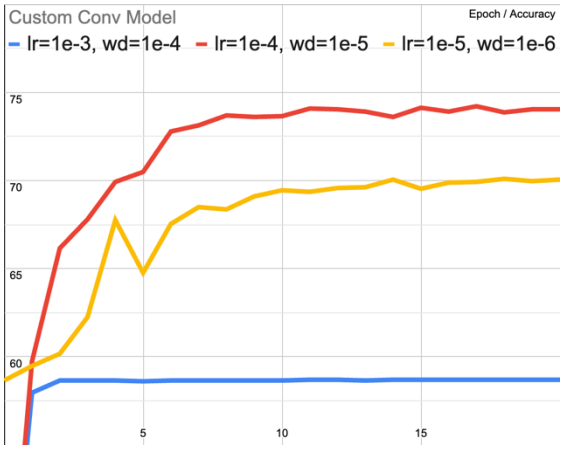

Figure 3. Custom model being trained on corpus of viridis spectrograms. With best performance of 74% (red line)

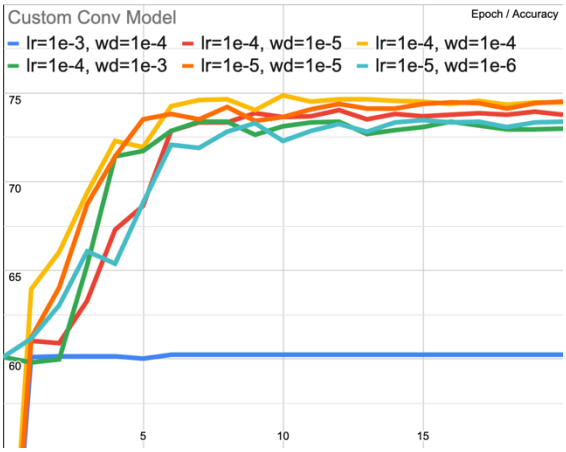

Figure 4. Custom model being trained on corpus of grayscale spectrograms. With best performance of 74.51% (orange line)

distinct forms of audio representation. The outputs from these feature layers are then concatenated and fed into a classifier layer.

The addition of the Chromagram-specific feature head enables the model to capture audio characteristics, such as pitch, which are not represented in Mel spectrograms. This addition enhances the model's capability to discern gunshot sounds with greater accuracy.

B.4 Vision Transformer

We included the Vision Transformer (ViT)[4], another state-of-the-art model to benchmark against our custom CNN and EffNetV2 models. We wanted to establish a point of reference for the effectiveness of our custom-designed models. Training and assessing the ViT on our dataset is expected to yield insights that could inform potential improvements in our model's architecture. This comparative analysis is not only aimed at gauging the maximal or near-maximal accuracy attainable on our dataset but also at enhancing our understanding of the performance dynamics between CNNs and transformer-based models in spectrogram analysis.

C. Experiments

C.1 Custom CNN Model on Viridis Dataset

The objective of this experiment was to optimize the performance of our custom CNN model using the viridis dataset. By adjusting the learning rate and weight decay hyper-parameters through a range of values, we aimed to identify the optimal configuration that yields the highest accuracy (see Figure 3 & Table 1).

Table 1: Configuration comparison on 20 epochs of training on the viridis color scheme corpus.

Learning Rate	Weight Decay	Accuracy
1e-3	1e-4	58.69%
1e-4	1e-5	**74.04%**
1e-5	1e-6	70.05%

C.2 Custom CNN Model on Grayscale Dataset

In line with past research[5], which suggests that utilizing grayscale images could significantly enhance accuracy in certain contexts, we explored the performance of our custom CNN model on the grayscale dataset. However, our experiments did not yield a significant improvement in detecting gunshots (see Figure 4 & Table 2).

Table 2: Configuration comparison on 20 epochs of training on the grayscale corpus of spectrograms.

Learning	Weight Decay	Accuracy
1e-3	1e-4	60.24%
1e-4	1e-5	73.77%
1e-4	1e-4	74.47%
1e-4	1e-3	72.99%
1e-5	1e-5	**74.51%**
1e-5	1e-6	73.38%

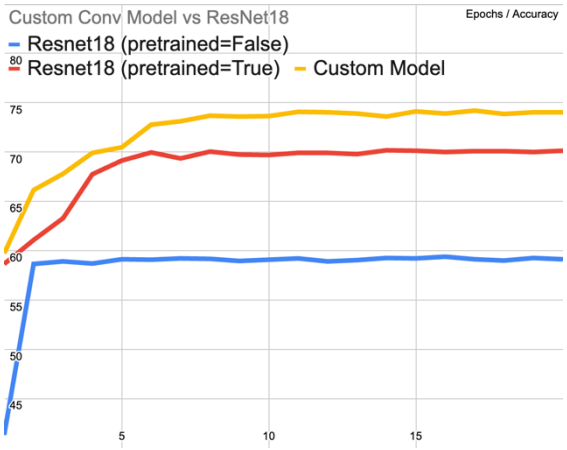

Figure 5. Comparison of accuracy between ResNet18 with no pretrained weights (blue), with pretrained weights (red) and Custom Model (yellow)

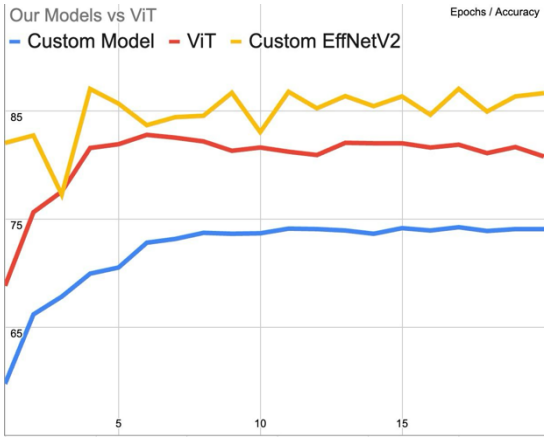

Figure 6. Comparison of accuracy between Custom CNN Model (blue), Custom Dual-headed EffNetV2 (yellow) and ViT (red).

C.3 Comparison of accuracy between Custom Model and ResNet18.

We trained the ResNet18 model using the same viridis dataset to benchmark its performance against our custom CNN model. While incorporating pre-trained weights significantly enhanced the accuracy of ResNet18, our analysis revealed that it still did not surpass the accuracy achieved by our custom model (See Figure 5 and Table 3).

Table 3: Comparison of ResNet18 with custom CNN.

Model	Pretrained	Accuracy
Custom CNN	False	**74.04%**
ResNet18	False	59.04%
ResNet18	True	70.22%

C.4 Feature selection for dual-headed EffNetV2 and Maximum Achievable Accuracy.

In response to the accuracy levels achieved in previous experiments, we explored other architectures that may potentially improve our results. Recognizing the ability of EffNetV2 in processing complex patterns and diverse data, we customized it with 2 feature heads instead of 1, to process two image sources.

The primary challenge was determining the optimal type of data to feed into the model's second feature head, with the first already dedicated to Mel spectrograms.

After analyzing the information encapsulated by Mel spectrograms and identifying what elements were missing, we selected three potential features for the second feature head:

1) Phase
2) Chromagram
3) Short-Time Fourier Transform (STFT)

We conducted experiments to evaluate which feature addition would most effectively complement the Mel spectrograms. Our results (See Table 4) demonstrate that the addition of Chromagrams provided the most significant enhancement in accuracy compared to Phase and STFT.

Table 4: Comparison of audio features.

Feature	Accuracy
Phase	81.19%
STFT	85.56%
Chromagram	**86.73%**

In an effort to gauge the maximum or near-maximum achievable accuracy on our synthetic dataset, we considered the Vision Transformer (ViT) as a benchmark. While aware of the inherent challenges associated with our dataset's size – given that ViT models generally require large amounts of training data – we proceeded with the experiment. The ViT model did outperform our Custom CNN, albeit falling short of the results achieved by our Dual-Headed EffNetV2 model (See Figure 6 & Table 5).

Table 5: Comparison of Custom Model, ViT and Custom EffNetV2.

Model	Accuracy
Custom CNN	74.04%
ViT	82.74%
Custom EffNetV2	**86.73%**

III. Future Work

Our experiments underscored the significant accuracy gains achievable through dual-feature head models that simultaneously process two different spectrogram types or audio features. However, further exploration of the type of feature to augment the Mel spectrogram that most effectively enhances gunshot detection needs to be done.

In addition, the practical deployment of these models in a real-time detection environment presents challenges, primarily due to the models' sizes and the inference times involved. Future research will need to focus on optimizing the model architecture to be more efficient without compromising accuracy. This optimization is crucial for enabling real-world applications.

IV. Conclusion

Our study reveals the superior performance of the custom Dual-Headed EffnetV2 over ViT and custom CNN. Our research significantly contributes to audio-based surveillance, presenting a robust solution for automated gunshot detection using computer vision techniques. Despite expectations, the finetuned ResNet18 model did not outperform our custom CNN model, emphasizing the challenges of transferring models to tasks with unique spectrogram characteristics. While ViT faced challenges with our limited dataset, EffNetV2's impressive adaptability with diverse datasets highlights its efficiency in handling complex patterns.

References

[1] Incze, Á.; Jancsó, H.B.; Szilagyi, Z.; Farkas, A.; Sulyok, C. Bird sound recognition using a convolutional neural network. In Proceedings of the IEEE 16th International Symposium on Intelligent Systems and Informatics, Subotica, Serbia, 13–15 September 2018

[2] J. Allen, "Short term spectral analysis synthesis and modification by discrete fourier transform", IEEE Transactions on Acoustics Speech and Signal Processing, vol. 25, no. 3, pp. 235-238, Jun 1977.

[3] Tan, Mingxing, and Quoc Le. "Efficientnetv2: Smaller models and faster training." International conference on machine learning. PMLR, 2021.

[4] Dosovitskiy, Alexey, et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." arXiv.org, Oct. 22, 2020. Accessed Dec. 14, 2023.

[5] Xie, Yiting, and David Richmond. "Pre-training on Grayscale ImageNet Improves Medical Image Classification." ECCV Workshops, 8 Sept. 2018.