Scalable Gunshot Detection Systems with Convolutional Neural Networks

Alex Morehead[1], Lauren Ogden[2], Gabe Magee[3], Ryan Hosler[4], Dr. George Mohler[4]

[1]Department of Computer Science, Mathematics, & Physics, Missouri Western State University; [2]Department of Computer Science, Columbia University;
[3]Department of Computer Science, Pomona College; [4]Department of Computer and Information Science, IUPUI School of Science

Introduction

Many cities with gunshot detection systems depend on expensive systems that rely on humans differentiating between gunshots and non-gunshots, such as ShotSpotter®. Thus, a scalable gunshot detection system that is low in cost and high in accuracy would be advantageous for a variety of cities across the globe, in that it would favorably promote the delegation of tasks typically worked by humans to machines.

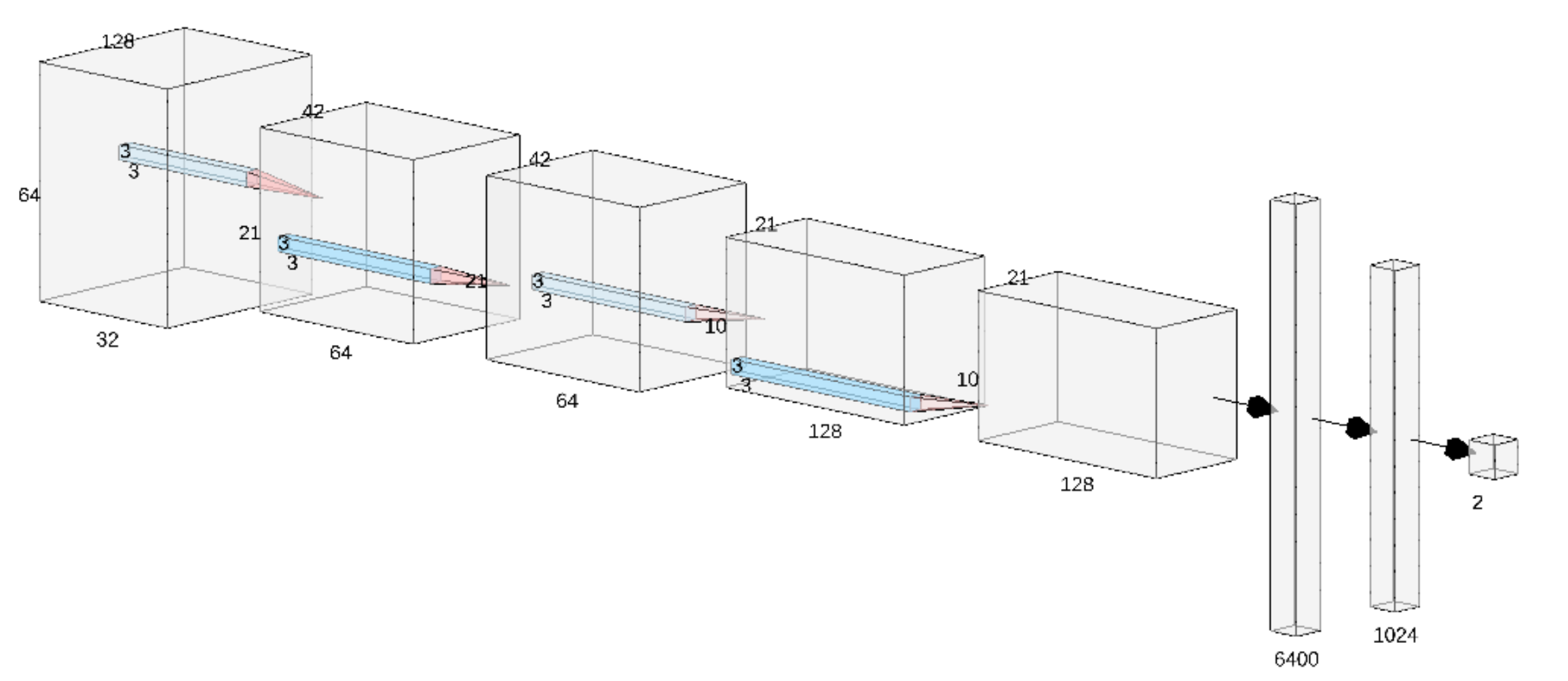
The Data

We obtained our data from two places: free internet databases such as Freesound and a repository of sounds recorded using a microphone connected to a Raspberry Pi microcomputer. In addition to this, we used a generative adversarial network (GAN) as well as sound augmentations to create additional samples of gunfire sounds and to prevent our model from overfitting to our compiled dataset.

Time Shift	Pitch Change	Speed Change	Volume Change	Background Noise Addition
Shifts a sound sample to the left or right by a randomly chosen amount less than 50% of the length, and then fills in silence as needed.	Changes the pitch of a sample by a randomly-chosen factor between 70% and 130%.	Alters the playback speed of a sample by a randomly-chosen amount between 70% and 130%.	Increases the amplitude of a sample with a uniformly-random variable.	Introduces random background noise into a sample while making sure that no gunshots are added into a sample that does not originally contain a gunshot.

CNNs

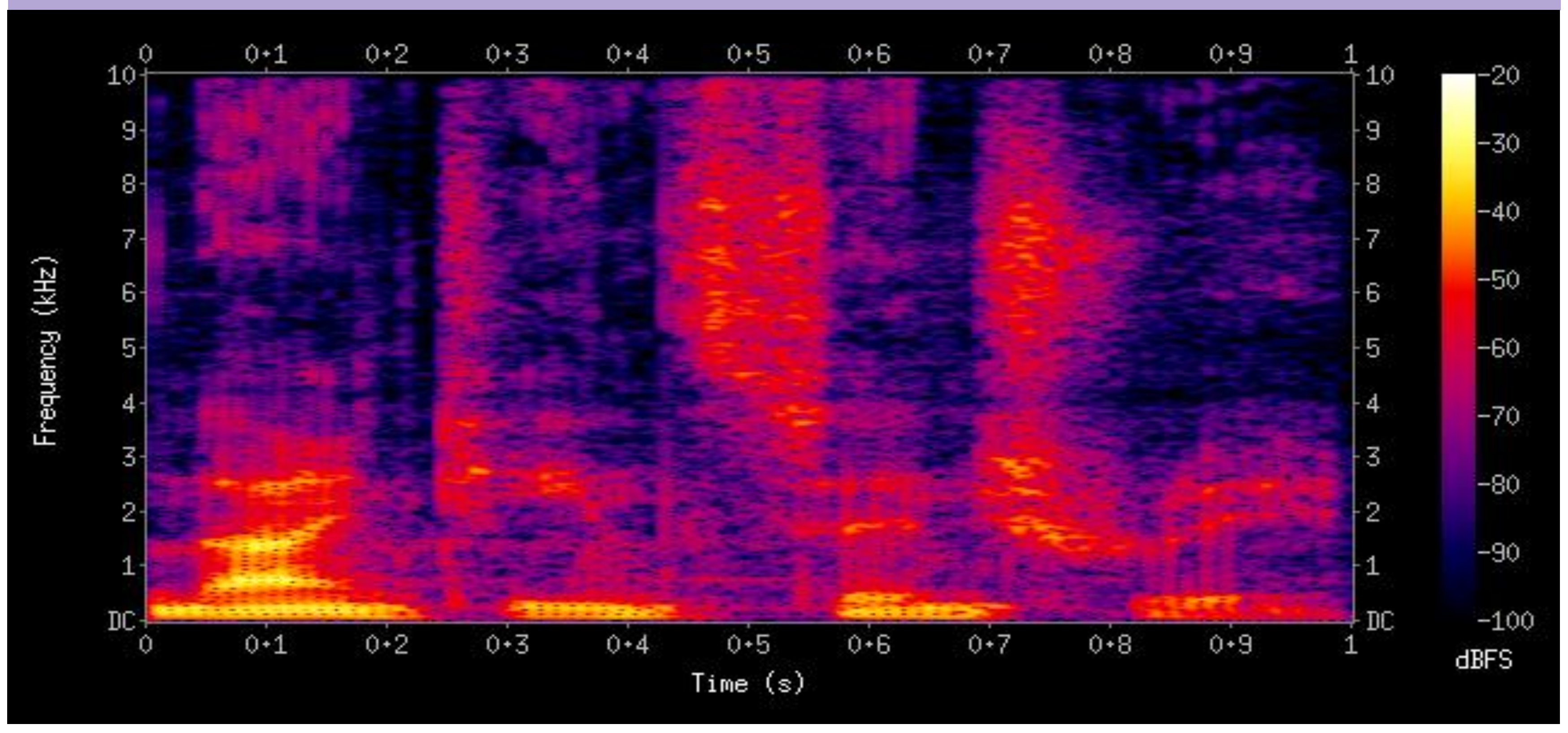
Convolutional Neural Networks (CNNs) are neural networks designed to locate, model, and accurately predict patterns present in input data such as a colored image.



They do so by iteratively sliding over small regions of data and translating any inherent properties in a region over to a proceeding network layer. This process is repeated up until the output layer which generates a prediction.

Spectrograms

Spectrograms are visual representations of the frequency and amplitude of sound over a specified span of time. For our project, we created two models, a 1D architecture that looks at sound represented as an array of frequency values and a 2D architecture that instead analyzes sound represented as spectrograms. For the time-series model each entry simply corresponds to a frequency measurement in a time-series, whereas for the spectrogram model the entry for each specific frequency-time cartesian coordinate is an amplitude value.



Methodology

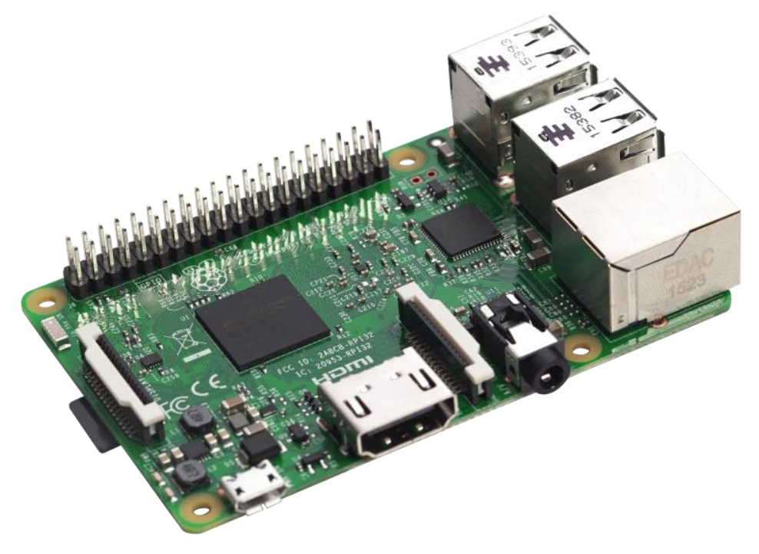
Development Process

Training

A convolutional neural network (CNN) was trained on a variety of sound data to recognize gunshots. While we had labels for sounds other than gunshots, we grouped them into a singular group "other". Then each of these samples were preprocessed to turn them into spectrograms if the model was trained on them. The models were trained for 100 epochs or until the target metric of a training session, accuracy in this case, did not change for fifteen epochs.

Deployment

Each model was then deployed to a Raspberry Pi Model 3 B+ with an SMS modem attached. In order to conserve space, we decided to convert them to TensorFlow Lite (TFLite) models, with a tradeoff for performance time. Our program has three processes – one to put audio from a stream onto a queue, one to analyze sound data pulled from the queue, and one to send an SMS alert message to a predetermined list of phone numbers.



Gunshot Detection Results

Using a validation data set, we found that our Keras models performed reasonably well with predictions. Regarding the performance of our models when stored in TFLite format, the results below describe expected levels of effectiveness.

	2D Convolutional Neural Network (128 x 64)		2D Convolutional Neural Network (128 x 128)		Convolutional Neural Network Ensemble (64 Model + 128 Model)	
	Keras	TFLite	Keras	TFLite	Keras	TFLite
Accuracy	98.8%	N/A	98.8%	N/A	98.8%	N/A
Precision	96.5%	N/A	94.3%	N/A	97.1%	N/A
Recall	93.8%	N/A	96.2%	N/A	92.9%	N/A
F1 Score	95.1%	N/A	95.2%	N/A	95.0%	N/A

Significance & Future Work

The findings generated by this research project have the potential to expand the current state of knowledge regarding sound-based applications of CNNs, and while simultaneously reducing the amount of jobs that require human input the results of this project could very well increase the standards of safety for a city's residents. Ideally, a feature we would like to implement in our pipeline in the future would allow for robust localization of gunshot alerts within a proposed cluster of Raspberry Pi units.

Acknowledgements

We gratefully acknowledge the support of NSF grant REU-1659488 which provided this project's research stipends, travel funds, and supply money.