Explainable Multi-Modal Threat Detection from Audio and Location Data

Harish Nandhan Shanmugam and Manidatta Anumandla

Course CSCI 5922, University of Colorado Boulder

Abstract. This projects present a multi modal deep learning approach designed to classify the audio events into multiclass risk levels such as Normal, Potential Threat and Danger based on the contextual infomation from both the sound and location. Unlike the traditional audio classification systems that operate in isolation, here this approach integrates the location metadata to better understand the situational relevance of an audio cue. This approach uses a transformer based architecture for classification and propose an additional NLP explanation generator at enhance the interpretability. This privacy preserving system avoids the reliance on visual surveillance, making it ideal for the sensitive environments like schools, hospitals and public spaces. This work aims to bridge the gap between the accurate threat detection and actionable human readable insights which supports real-time and ethical deployment in the public safety applications.

Keywords: Audio Classification, Risk Detection, Deep Learning, Transformer, Context-Aware AI, Privacy, Multi-Modal Model, Explainable AI

1 Introduction

In the current era of heightened concern about the public safety and data privacy, having the ability to detect the threats in real time is more important than ever. Traditional surveillance systems does often relies in the video feeds, raising the ethical and privacy concerns. Our world, is richer with sound cues which can convey critical situational awareness which includes gunshots in a school, alarms in the bank, crowds in the protest etc. Automatically identifying the risk level with an audio event, which is contextualized by its location does offer a powerful tool for the smart, privacy-preserving safety systems. This approach will not only reduces the need for intrusive video surveillance bust also makes the threat detection with more proactive and context-aware.

Existing audio classification models has the ability to detect the sounds such as "gunshot" or "siren" but they often do so in the isolation which fails to take where the sound has actually occurred. A scream at the convert may not indicate the concern but the same sound in the hospital corridor sound indicate the distress. Addition to it, the current models suffer from the lack of transparency, they doesn't provide reasoning behind the classifications, it may predict the sound level but will not tell about why it is actually occurred, which is unacceptable in the high-stakes in the domains like the emergency response or airport security. Furthermore, the most systems treat the risk as the binary or overly simplistic, ignoring the nuanced contextual understanding.

These challenges highlighting the need for systems which not only detect sound events but also to interpret them within the real-world context. Having effective threat monitoring system must go beyond the raw classification to the incorporate environmental cues like location and provide the interpretable outputs which aligns with the human understanding. These systems should be capable to differentiate the identical sounds which imply different levels of risk depending on where and when they will occur. Furthermore, to ensure the usability in mission critical applications, the system should offer the explanations for its predictions to build the trust among operators and support the timely, informed decisions. Addressing these needs will require shift towards the context-aware, explainable and the ethically designed models.

We will propose the two-stage deep learning system for our audio based threat monitoring:

1. A transformer-based classifier which fuses the audio and location data to predict the risk level into three different categories Normal, Potential and Danger.

2. A generative model which produces the natural-language explanations to justify the classification based on the both input and the predicted risk.

This novel system is multi-modal, interpretable and the privacy-preserving. It introduces the contextual awareness and has the human centered design to the traditional audio classification. By doing thus, it will fill the gap on public safety monitoring systems which should balance the real-time detection and the ethical deployment.

Our model outputs the clear, categorical risk levels which are easy to understand and act upon. The focus on the interpretable classifications better integrates with the alert systems, dashboards and also human operations which needs the actionable insights in the real time. The audio-locations fusion model promotes the deployment in camera-free environments which enhances the privacy.

2 Related Work

2.1 Audio Event Classification

Recently developed models in audio classification through large scale datasets and deep learning models [1] introduced the Audioset, an ontology and dataset of audio events. Most recent work by the [2] reviwed the deep learning methods in audio classification which highlights the role of CNNs, RNNs and hybrid approaches. [3] proposed CNN architectures for robust audio classification. [4] developed FSD50K, a comprehensive dataset with human labeled audio events.

[5] has provided the in-depth overview of deep learning strategies which are applied to broad range of audio signal processing tasks. Our approach will extend this by contextualizing the risk based on including location metadata, addressing practical limitations in high-stakes safety applications.

2.2 Context-Aware Classification

[6] introduced context - aware models for the surveillance using the environmental sound. [7] demonstrated multi-modal learning combining the sound and scene context. This Prior work has introduced models capable of contextual sound understanding, such as environment surveillance systems incorporating the ambient context. Multi-modal learning has been emerged, combining the auditory input with the visual data enhancing the model accuracy. Our technique advances these efforts by employing the transformer based model to fuse the audio and location data, shifting focus from general scene recognition to the risk-level prediction, which is a critical factor in real-time safety monitoring systems.

2.3 Explainable AI in Audio

[8] presented an audio captioning dataset and baseline, while [9] explored the sound event detection with the synthetic data. These studies has explored the audio captioning and sound event detection using the synthetic data, offering the early methods for interpretability. Our model will contribute the explanation-generation component, which produces the natural language justifications which are tailored to the predicted risk level and the location. This features ensures that system's outputs are transparent and actionable for the human operators.

3 Methodology

3.1 Risk Level Classifier

- **Input:** A 5-second audio clip which is transformed to a log mel spectrogram and its associated encoded location metadata.
- **Audio Processing:** We used Librosa for the coversion of raw audio files into 128 bin log-mel spectrograms with a sample rate of 22,050 Hz. All the clips were standardized to a 5-second clip. The short clips were padded with zeros. Spectrograms were normalized and resized to a fixed shape of 128 x 216 to allow consistent input dimensions.
- **Architecture:** A Convolutional layer segments the 128x216 spectrogram into a non-overlapping 16x16 patches. These patches are then flattened and passed into the Transformer Encoder with the positional embeddings to capture the temporal and spectral dependencies. The location metadata is embedded into the learnable vector and concatenated with the audio representation.

- **Output:** A fully connected feed forward neural network predicts one of three risk levels into Normal, Potential Threat and Danger.
- **Dataset:** FSD50K which contains the audios of different living and non-living things, augmented with the location metadata and the risk levels.
- **Training & Optimization:**Hyper parameter tuning was conducted using Optuna, exploring the combinations of learning rate, hidden dimension size, attention heads, batch size and the optimizer. The best configuration is with the learning rate of 0.00327, hidden dimension of 128, heads of 4, batch size of 32 and Adam optimizer.
- **Design Rationale:** Mel-spectrograms preserve temporal and frequency content which is making them ideal for capturing nuances in environmental audio. Adding the location metadata disambiguates acoustic events with similar spectral patterns but different level of contextual implications. For example, Sirens at home vs Public places.

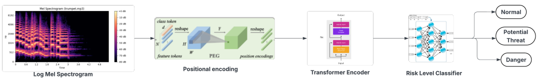

Fig. 1. Risk Level Classifier

3.2 Explanation Generator

- **Input:** The predicted risk level from the risk level classifier model, label example: "Gunshot" and location example: "School".
- **Architecture:** A hugging face T5 Small model was fine-tuned on custom template based training data. The input format here should be "Audio, Location and Risk" and the output format should be One of 15 diverse natural language rationales per input. The decoder generates the fluently worded statements like "A dangerous sound was detected in a school. Immediate attention is required."
- **Training & Evaluation:** Cross Entropy loss was used for training. BERTScore was used to assess the semantic similarity of the generated outputs vs reference texts. The model achieved a BERTScore F1 of 98
- **Design Rationale:** Explanation generation bridges the gap between AI predictions and human interpretability. Custom templated data helped the model learn diverse phrasing for the same intent. The T5 architecture enables the transfer learning and fluent language generation, making it suitable for the safety critical applications.

float

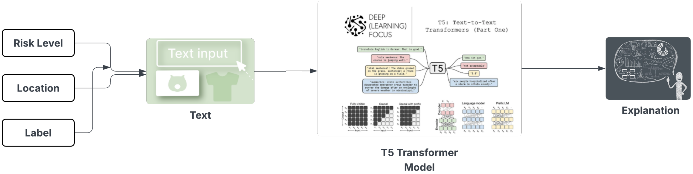

Fig. 2. Explanation Generator

4 Experiments

4.1 Evaluation of Risk Level Classifier

– **Objective:** Validating the effectiveness of the audio location fusion using Transformer architecture for the risk classification.
– **Training Strategy:** Hyperparameter tuning is done using Optuna over 10 trials. Key parameters explored are Learning rate, Hidden dimension, No. of attention Heads, Optimizer and Batch Size. The best configuration is with the learning rate of 0.00327, hidden dimension of 128, heads of 4, batch size of 32 and Adam optimizer.
– **Performance Metrics:** The Validation Accuracy of 87.4% was achieved on the best trial i.e Trial 3. A test accuracy of 87.2% was achieved and 'Adam' optimizer with moderate hidden dimension performed consistently well but the performance plateaued after 5 epochs.

4.2 Evaluation of Explanation Generator

– **Objective:** Assess the model's ability to produce the coherent and relevant explanations.
– **Training Strategy:** Finetuned the Hugging Face T5-Small model using input-output text pairs and evaluated it using the BERTScore for semantic for semantic similarity with reference texts.
– **Performance Metrics:**

Epoch	Training Loss	Validation Loss	Precision	Recall	F1 Score
1	0.3229	0.2075	0.9473	0.9504	0.9488
2	0.1907	0.1075	0.9720	0.9739	0.9729
3	0.1358	0.0816	0.9756	0.9773	0.9765
4	0.1205	0.0715	0.9791	0.9801	0.9796
5	0.1176	0.0691	0.9795	0.9806	0.9800

Table 1. Results for Explanation Generator

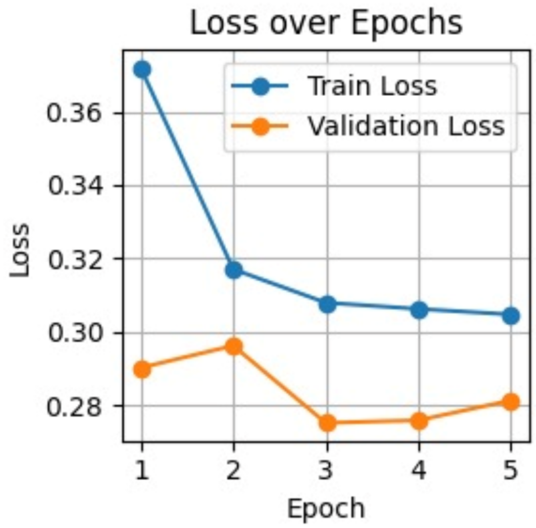

Fig. 3. Loss vs Epochs for Risk level Classifier

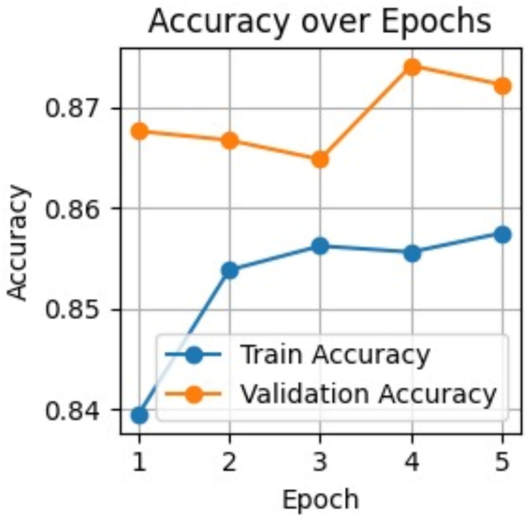

Fig. 4. Accuracy vs Epochs for Risk level Classifier

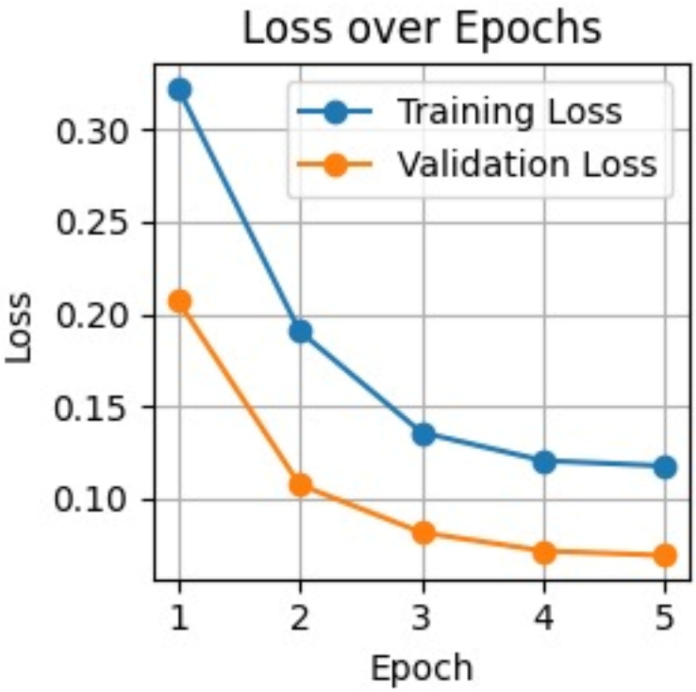

Fig. 5. Loss vs Epochs for Explanation Generator

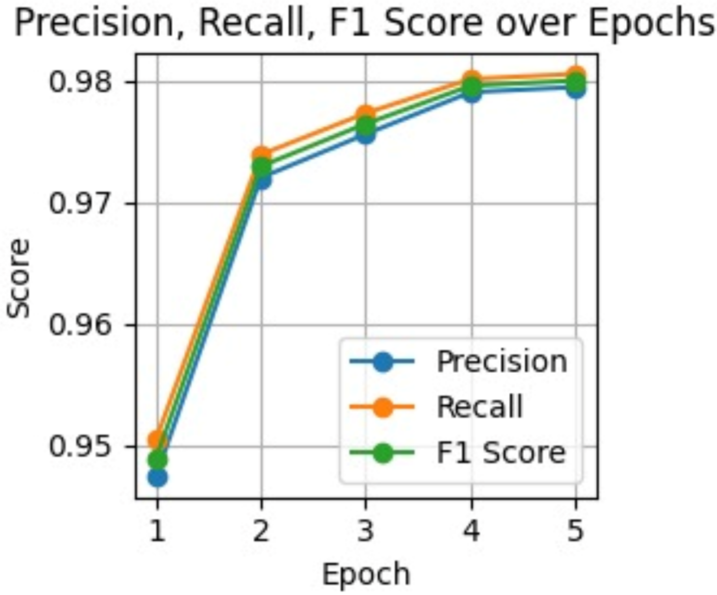

Fig. 6. Precision, Recall, F1 Score for Explanation Generator

4.3 Open Questions & Next Steps

- **Data Limitation:** Assuming the label and location are known at inference. Future work could explore the models that extract all attributes (label, location,risk) directly from raw audio.
- **Temporal Dynamics:** Current model operates on static 5-second clips. Exploring sequential audio processing (via LSTMs or audio transformers) could help detect evolving threats.
- **Explanations for Ambiguity:** For borderline cases (eg. Potential Threat vs Danger), the explainability generator could integrate uncertainty scores or generate multi-hypothesis justifications.
- **Bias Auditing:**Additional analysis is needed to check if location or label biases affect classification or explanation reliability.

5 Conclusions

In this project we are presenting a novel approach of context aware risk classification framework that uses audio and location information to improve the public safety in a privacy preserving manner. Our approach not only detects the risk levels from the audio cues using a transformer based classifier but also offers a human readable explanations to improve the interpretability and trust in real world applications. By avoiding the intrusive surveillance methods like video and emphasizing the transparency through explainable AI, our system addresses the critical challenges in the modern thread detection. The inclusion of multiple experiments including the accuracy evaluation and explanation generation, ensures comprehensive validation. This work lays the foundation for scalable, ethical and effective deployment of AI in smart security systems. The proposed system demonstrates strong performance (87.2% classification accuracy and 98% BERTScore F1) on real-world, context-augmented environmental audio.

6 Ethical Considerations

By generating the explanations, we address the black box nature of the traditional models and improve interpretability. The use of location metadata must be handled with caution. Anonymization or any form of user consent protocols are crucial for the deployment. Further fairness checks are required to ensure that the model does not disproportionately assign higher risk levels based on the location patterns, which could mirror socioeconomic biases. This system can enhance the emergency response, especially in high risk environments (Eg.Schools,Malls,Transport hubs) but misuse (Eg: Surveillance or biased threat escalation) must be mitigated via policy and human oversight.

References

1. Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE (2017) 776–780
2. Bose, A., Tripathy, B.: Deep learning for audio signal classification. Deep learning research and applications (2020) 105–136
3. Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., et al.: Cnn architectures for large-scale audio classification. In: 2017 ieee international conference on acoustics, speech and signal processing (icassp), IEEE (2017) 131–135
4. Fonseca, E., Favory, X., Pons, J., Font, F., Serra, X.: Fsd50k: an open dataset of human-labeled sound events. IEEE/ACM Transactions on Audio, Speech, and Language Processing **30** (2021) 829–852
5. Purwins, H., Li, B., Virtanen, T., Schlüter, J., Chang, S.Y., Sainath, T.: Deep learning for audio signal processing. IEEE Journal of Selected Topics in Signal Processing **13**(2) (2019) 206–219
6. Soni, S., Dey, S., Manikandan, M.S.: Automatic audio event recognition schemes for context-aware audio computing devices. In: 2019 Seventh International Conference on Digital Information Processing and Communications (ICDIPC), IEEE (2019) 23–28
7. Liang, H., Ji, W., Wang, R., Ma, Y., Chen, J., Chen, M.: A scene-dependent sound event detection approach using multi-task learning. IEEE Sensors Journal **22**(18) (2021) 17483–17489
8. Mei, X., Meng, C., Liu, H., Kong, Q., Ko, T., Zhao, C., Plumbley, M.D., Zou, Y., Wang, W.: Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. IEEE/ACM Transactions on Audio, Speech, and Language Processing (2024)
9. Serizel, R., Turpault, N., Shah, A., Salamon, J.: Sound event detection in synthetic domestic environments. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE (2020) 86–90