FRAMEWORK FOR EVALUATION OF SOUND EVENT DETECTION IN WEB VIDEOS

Rohan Badlani*, Ankit Shah*, Benjamin Elizalde, Anurag Kumar, Bhiksha Raj

Language Technologies Institute, School of Computer Science, Carnegie Mellon University

rohan.badlani@gmail.com, aps1@andrew.cmu.edu

Carnegie Mellon University
Language Technologies Institute

Introduction

- Lack of Annotated data poses a problem for Large scale learning of audio events.
- Audio data from web is unexplored since videos have no tags or labels for sounds at segment level.
- Introduce a framework for Large scale audio event recognition on web videos
- Explore the extent to which Search query can estimate audio event recognition system performance.

Proposed Framework

- **Crawl** Downloads YouTube videos using Pafy API.
- **Hear** Consists a Dataset Aggregator, Feature Extractor and Sound event classifier. Web audio is fed to Hear module for preprocessing and prediction of sound events.
- **Feedback** Displays the classifier prediction on the website nels.cs.cmu.edu for evaluation based on user feed back.

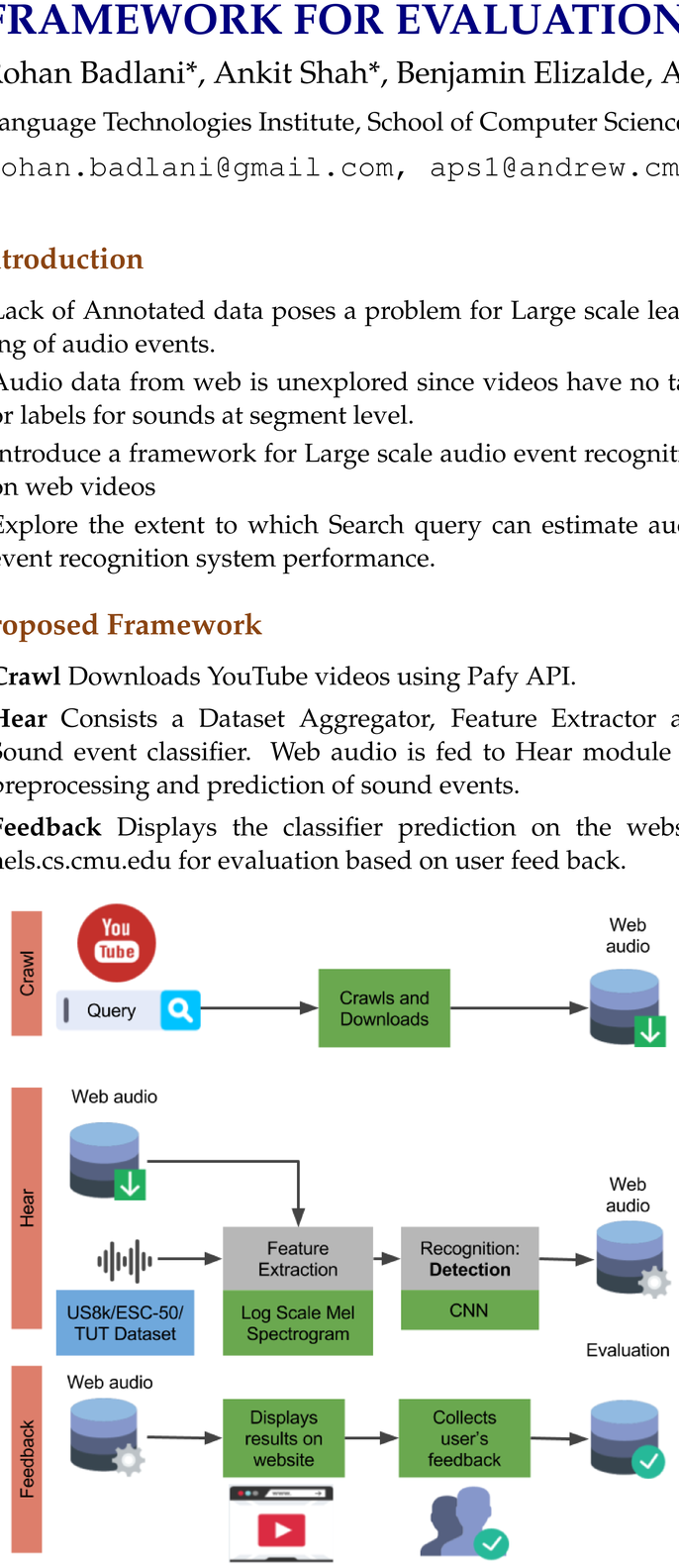


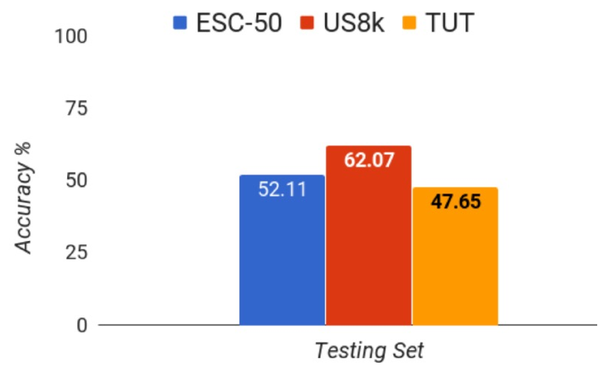
Figure 1: Framework consists of three modules: Crawl, Hear and Feedback

Experiments and Results

- **Data** - **78** Sound Events from 3 datasets: ESC-50, TUT and UrbanSound8k. Our crawler module downloaded **7600** YouTube video's soundtrack (**260** hours of audio or 3.5 million audio segments) with 100 audio recordings for each category.
- **Features** - Log scaled mel spectrograms with 60 mel bands, a window size 23ms and hop size 11.5 ms from 16 bit, mono channel audio at 44.1kHz resampled audio.
- **Classifiers** - Convolutional Neural Network [1] (one for each dataset) were trained to obtain prediction. Datasets were partitioned into 60% training, 20% testing and 20% validation sets

Text metadata from YouTube suggests sound presence at video level, thus no groundtruth is available at segment level to evaluate web audio.

- **Experiment 1** - Query reflects accumulated text like title, keywords and description. Thus, we use Query as groundtruth which means all the segments of a soundtrack have true class-label as corresponding search query used for retrieval.
- **Experiment 2** - Human labeling ⇒ reliable source of Ground truth, authors inspected the top 40 (K = 40, K is number of segments) segments based on classifier confidence and evaluated the classifier prediction providing feedback using website.



Performance trend for query and human feedback as groundtruth

- Similar Performance trend for query and human feedback with less than 10% difference in precision observed.
- Precision of Top K video segments (1_i=K_i=5) for the three classifiers is expected to be unstable. However, performance stabilizes as K grows.
- Precision trend of Top K video segments for the combination (weighted average) of the three trained classifiers suggests query is reliabile as the class label for sound events at segment level.

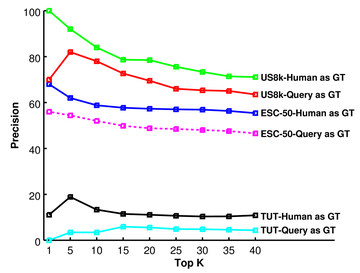


Figure 2: Search query-based performance similar trend to human feedback.

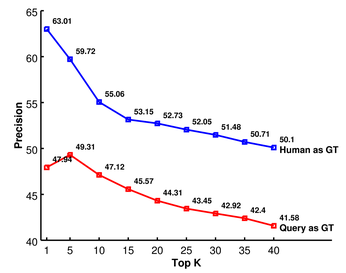


Figure 3: Performance for the combination of the three classifiers.

Conclusions

- Correlation between presence of audio events in video segments and the query used for its retrieval.
- Classifier prediction using search query and human feedback as groundtruth has similar trends showing correlation exists between sound events and queries in YouTube videos.
- Precision trend suggests that the search query could be a lower-bound of human inspection performance. Hence, the query could be used as class label at segment level reducing the dependency of annotated audio to estimate classifier performance.
- Our framework and its findings can be used to complement multimedia content retrieval algorithms, which are mainly based on text and images.

Future Work

Crowd Sourcing to inspect a larger number of segments.

References

[1] Karol J Piczak, "Environmental sound classification with convolutional neural networks," in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2015, pp. 1–6.