Knowledge Transfer From Weakly Labeled Audio Using Convolutional Neural Network For Sound Events and Scenes

Anurag Kumar∗, Maksim Khadkevich+, Christian Fugen +
alnu@cs.cmu.edu, khadkevich,fuegen@fb.com


In A Nutshell!
- CNN for large scale weak label learning
- Transfer learning using weakly labeled data
- State of Art Results on Audioset-Balanced Training
- State of Art Results on ESC-50 Dataset
- Outperforms human accuracy on ESC-50 dataset
- Establishing Semantic Relationships for Sounds

Weak Labels
- Acoustic Events and Scenes in audio recording
- Biggest Challenge - Labeled Data
- Annotations with time stamps of audio events

Figure 1: Strong Label vs Weak Labels

- Weakly Labeled Data: Presence or absence of events in the recording
- Weak Labels - Lesser labeling effort, can be automated

CNN for Weakly Labeled AED

General Idea - Work with Segments, Need to look through whole recording

Strong Label Assumption Training
- Ignore weak label - Assume event is present in whole recording
- Use your favorite CNN architecture

Prop: CNN for AED using Weak Labels
- Treat weak labels as weak
- Efficiently handle recording of variable length
- Single Forward pass - Computationally Efficient
- Segment and hop size controlled by network design

Transfer Knowledge
- Domain and Task Adaptation
  - Off the shelf representations
  - Adapt and then obtain representations
  - Train classifiers

Results - Weak Label Learning

Event Localization Examples in Audioset

Audio Event Classification on ESC-50

Audio and Video Forensics Conference (AVFS2016)

Results - Task Adaptation

Semantics for Sounds

Automatically Learn relationship between acoustic scenes and sound events

See which events from Audioset fires up for different acoustic scenes

Figure 2: Top Left and Right: Deep CNN for Weakly Labeled Audio. Bottom Left: Adapting CNN for target task. 3 different methods (I, II, III). Bottom Right: Obtaining representations for audio.

Figure 3: Left: Mean Avg. Precision and Mean Area Under ROC curves. Right: Computational Time Comparison.

Figure 4: Accuracy on ESC-50 dataset. ESC-50 - 50 Sound Events. Comparison with others in bar plot.

Figure 5: ESC-50 Accu Comparison with other methods.

Figure 6: t-SNE visualizations of learned representations. L: Higher 5 categories in ESC, R: DCASE16.