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

More details and Code - http://www.cs.cmu.edu/~alnu/TLWeak.htm

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

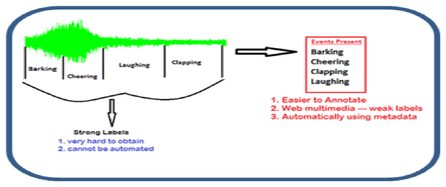


Figure 1: **Strong Label vs Weak Labels**

- 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

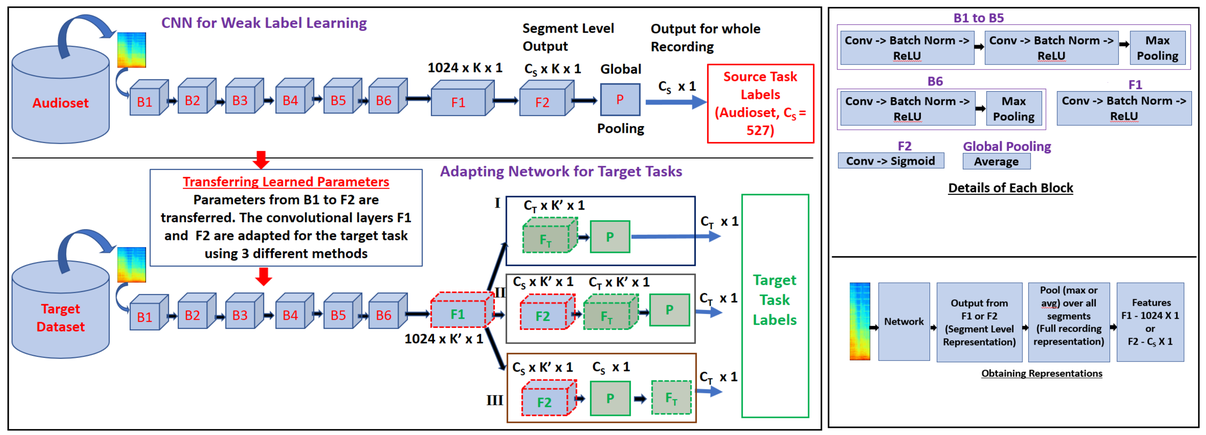


Figure 2: **Top Left and Right: Deep CNN for Weakly Labeled Audio. Bottom Left: Adapting CNN for target task. 3 different methods (I, II, III). Bottom Right: Obtaining representations for audios.**

Transfer Knowledge

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

Results - Weak Label Learning

Audioset

- Largest Dataset for sound events - Weakly Labeled
- 527 sound events
- *Balanced* and *Unbalanced* training set
- *Evaluation* set - $\sim$ 20,000 recordings

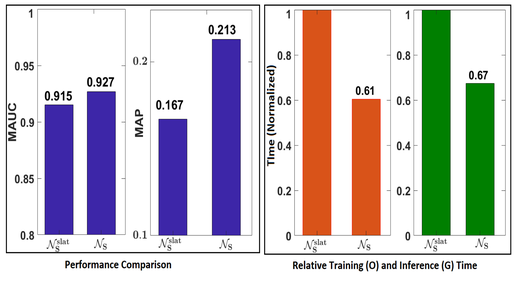
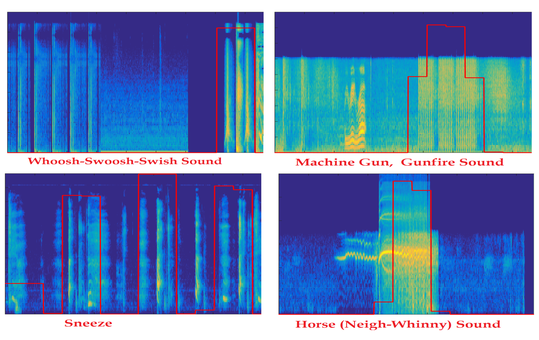


Figure 3: **Left: Mean Avg. Precision and Mean Area Under ROC curves, Right: Computational Time Comparison**

Event Localization Examples in Audioset



Results - Domain Adaptation

Audio Event Classification on ESC-50

Network	F1		F2	
	$max()$	$avg()$	$max()$	$avg()$
$\mathcal{N}_S$	**82.8**	81.6	65.5	64.8
$\mathcal{N}_T^I$	**83.5**	81.3	–	–
$\mathcal{N}_T^{II}$	**83.5**	81.8	81.9	81.5
$\mathcal{N}_T^{III}$	83.3	82.6	82.6	81.9

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

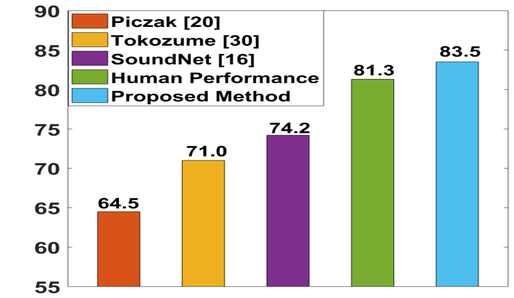
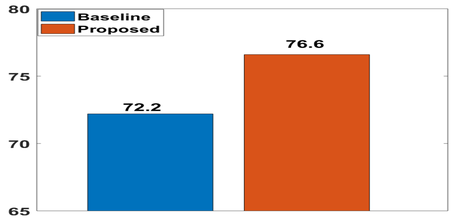


Figure 5: **ESC-50 Accu Comparison with other methods**

Results - Task Adaptation

Acoustic Scene Classification (DCASE16)



Semantics for Sounds

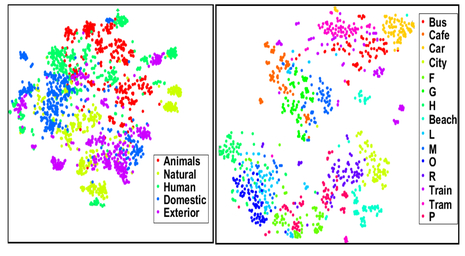


Figure 6: **t-SNE visualizations of learned representations. L: Higher 5 categories in ESC, R: DCASE16**

Automatically Learn relationship between acoustic scenes and sound events

See which events from Audioset fires up for different acoustic scenes

Scene	Frequent Highly Activated Sound Events
Cafe	Speech, Chuckle-Chortle, Snicker, Dishes, Television
City Center	Applause, Siren, Emergency Vehicle, Ambulance
Forest Path	Stream, Boat Water Vehicle, Squish, Clatter, Noise, Pour
Home	Speech, Finger Snapping, Scratch, Dishes, Baby Cry, Cutlery
Beach	Pour, Stream, Applause, Splash - Splatter, Gush
Library	Finger Snapping, Speech, Fart, Snort
Office	Finger Snapping, Snort, Cutlery, Speech, Cutlery
Residential Area	Applause, Crow, Clatter, Siren
Park	Bird Song, Crow, Stream, Wind Noise, Stream