AudioSet:
Real-world Audio Event Classification

g.co/audioset

Rif A. Saurous, Shawn Hershey, Dan Ellis, Aren Jansen
and the Google Sound Understanding Team
2017-10-20

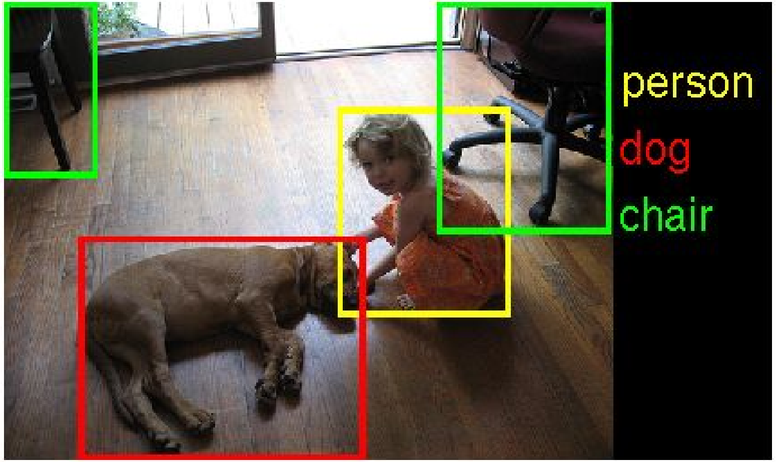
Research at Google

Outline

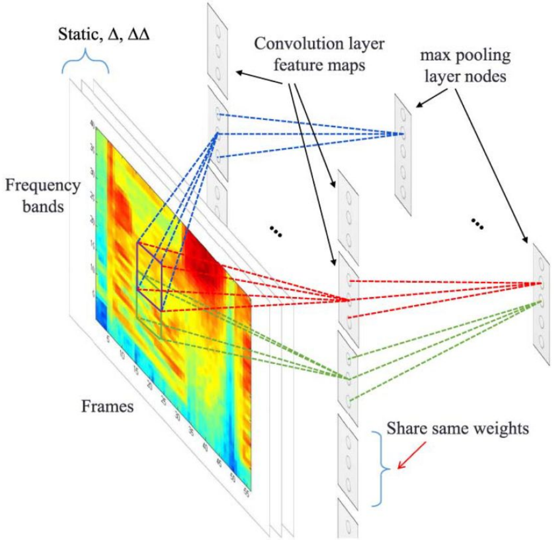
- The Early Years: Weakly-Supervised YouTube Videos
- AudioSet Is Born
- AudioSet: Supervised and Unsupervised

General Audio Event Classification

- Audio Event Classification
  Using ideas from:
  - ImageNet Object Recognizers



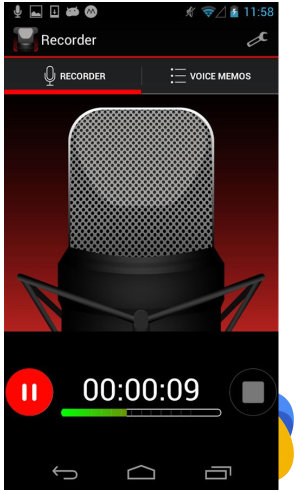
  - DNN Speech Recognition



Abdel-Hamid et al 2013

Audio Event Detection: Applications

- Content-based Archive Search
- Surveillance/Event Detection
- Context Awareness

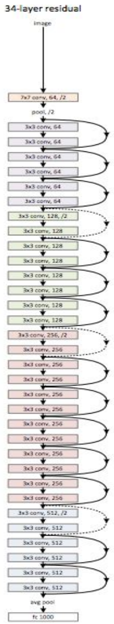


Web Video Classification - Summary

A TON of
Weakly Labeled
YouTube Audio

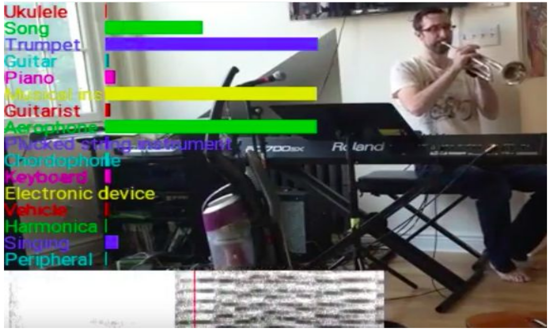
+

CNN architectures
from computer vision
community



=

Awesome Audio
Event Classification



Tasks of Interest

- Soundtrack Classification

- Audio Event Detection

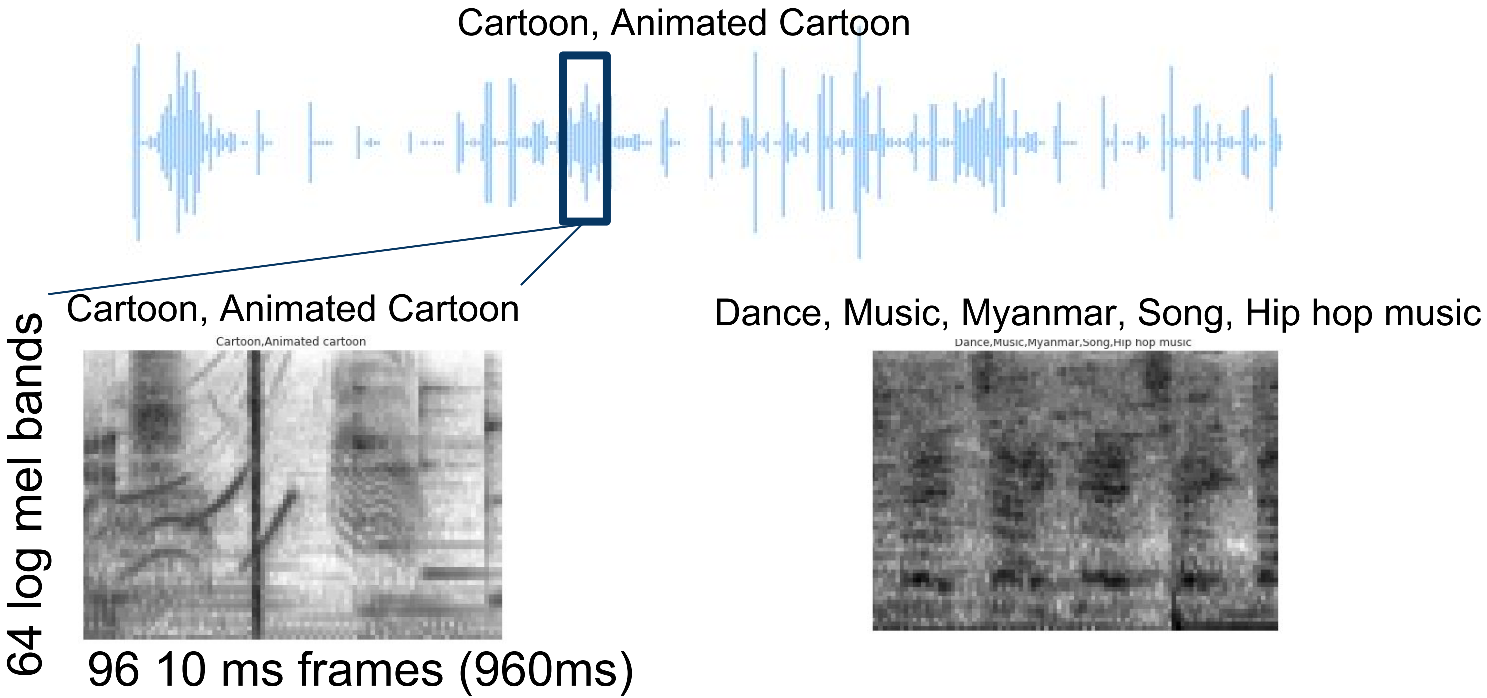
YouTube-100M DataSet

- Size
  - ~100M videos with video-level labels (~5 million hours, 600 years)
  - 20 billion input examples
- Labels
  - 30K labels (not all obviously acoustically relevant)
  - ~3 labels per videos

It's BIG

Label prior	Example Labels
$0.1 \ldots 0.2$	Song, Music, Game, Sports, Performance
$0.01 \ldots 0.1$	Singing, Car, Chordophone, Speech
$\sim 10^{-5}$	Custom Motorcycle, Retaining Wall
$\sim 10^{-6}$	Cormorant, Lecturer

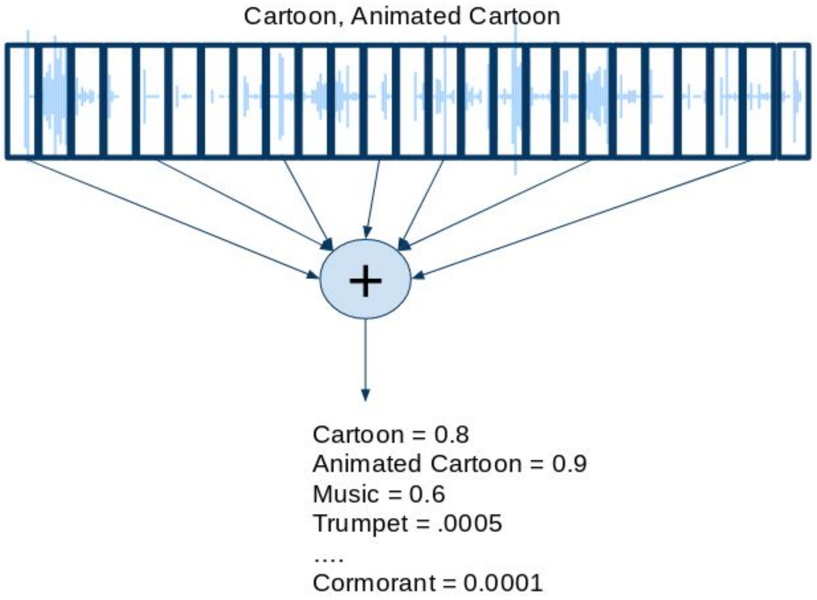
Training



- Train frame level classifier (very weak labeling)

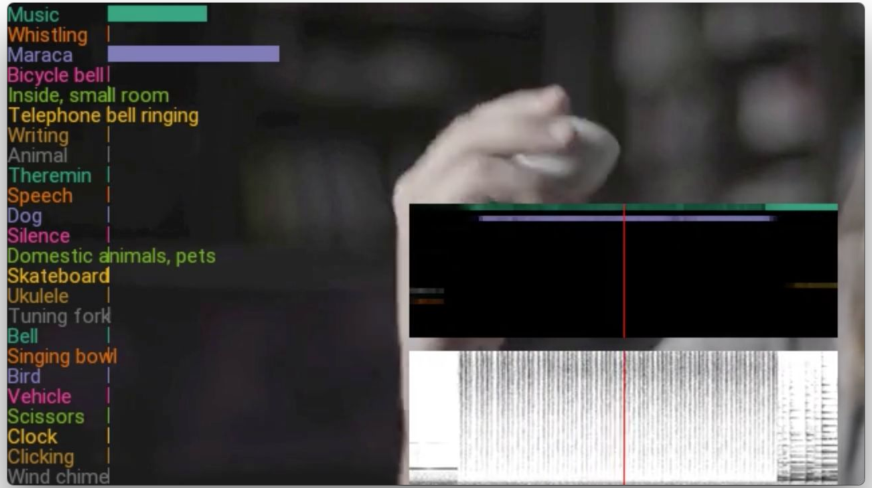
Evaluation

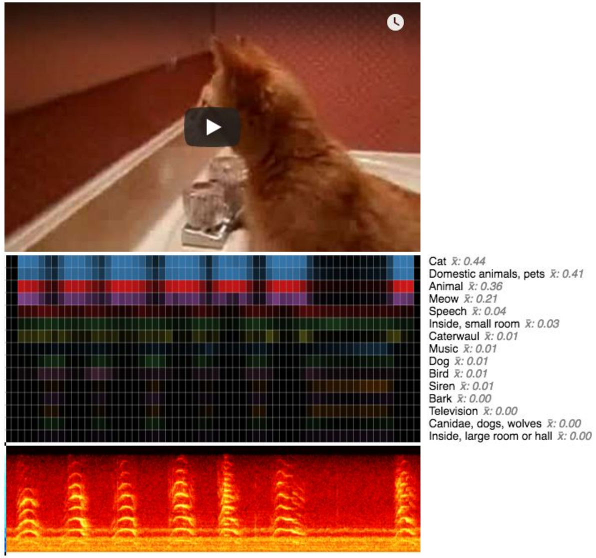
- Run frame-level classifier over each non-overlapping 960ms
- Aggregate over frames to evaluate video level scores
- Calculate mAP, AUC (DPrime)



Gut-feel Evaluation

- Look at ratings against a few favorite test cases
  - Potential problem: Focus on a few minor classes?

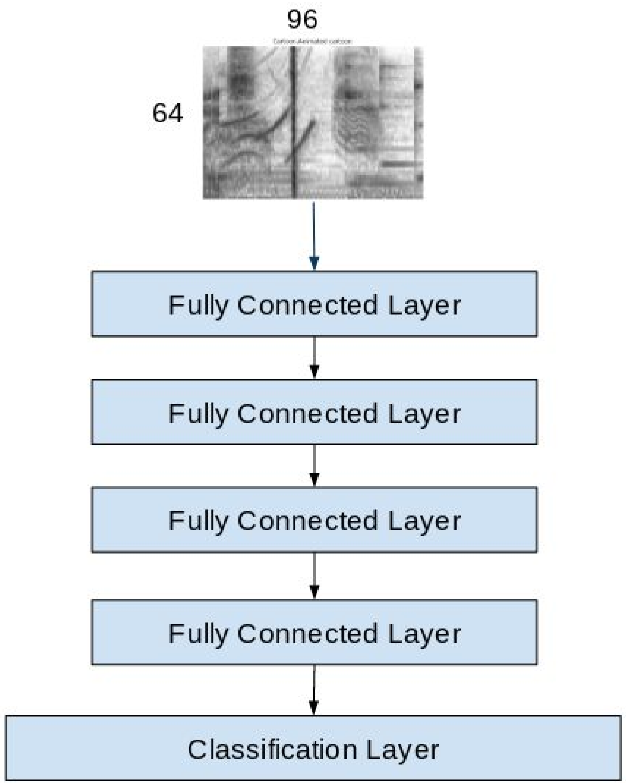


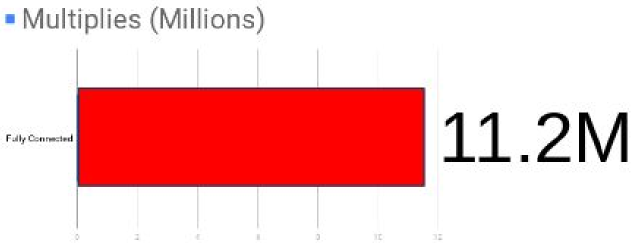


Questions

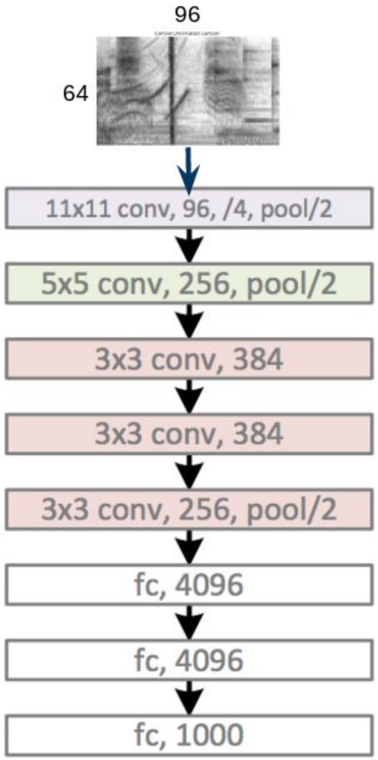
- Architectures - How well do various CNN architectures perform?
- Training Size - How do we benefit from training set size?
- Useful embeddings - Can we learn generally useful audio embeddings from our large dataset. (Embeddings that can be used as features to predict labels not in the original training set).

Architectures - Fully Connected





Architectures - AlexNet [Alex Krizhevsky et al. 2012]



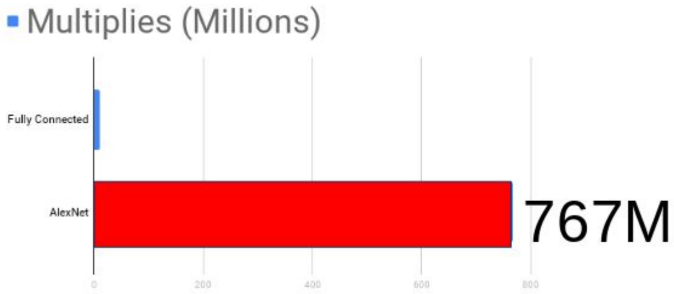
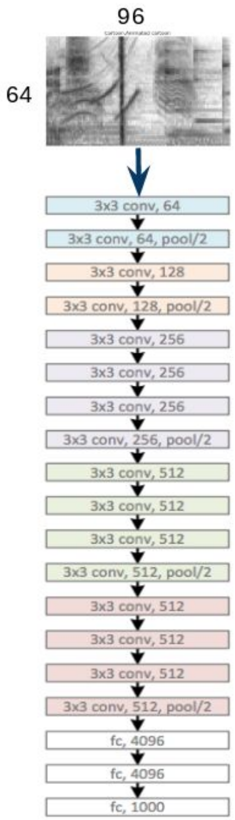


Image from Kaiming He's ResNet presentation

Architectures - VGG E [Karen Simonyan et al. 2015]



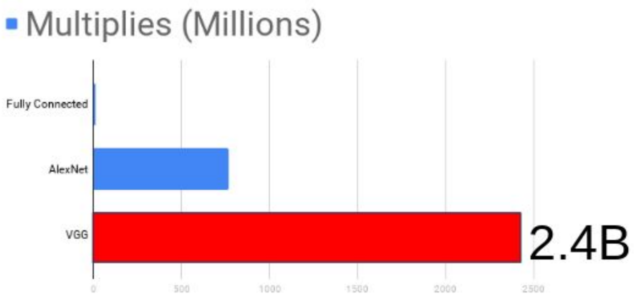
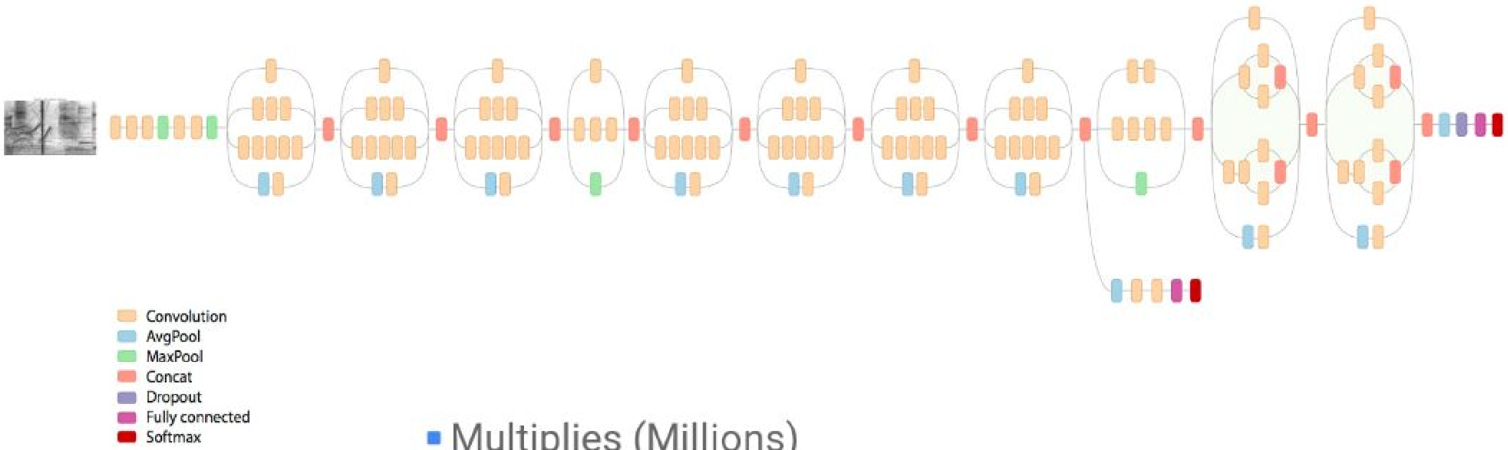
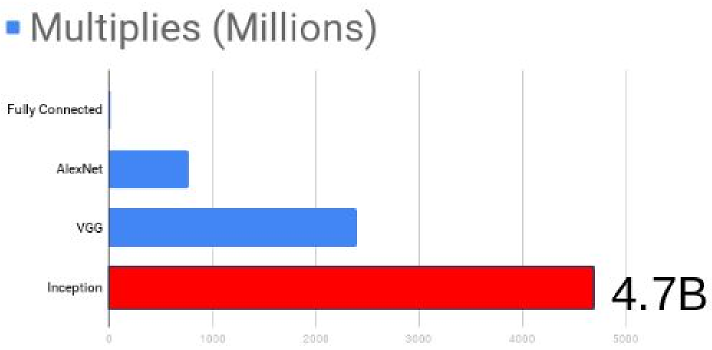


Image from Kaiming He's ResNet presentation

Architectures - Inception V3 [Christian Szegedy et al. 2015]





Architectures - ResNet [Kaiming He et al. 2015]

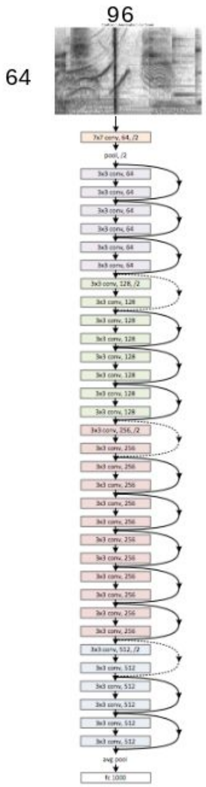
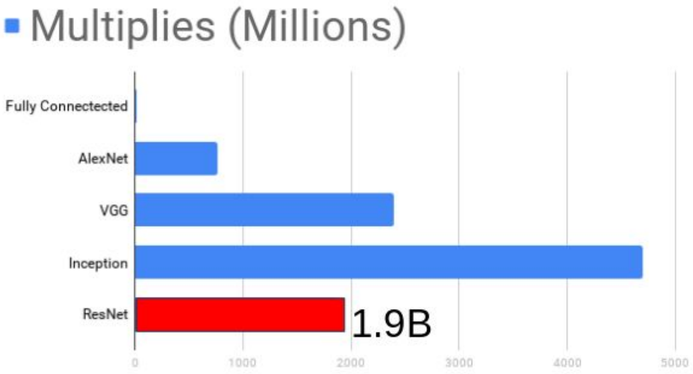
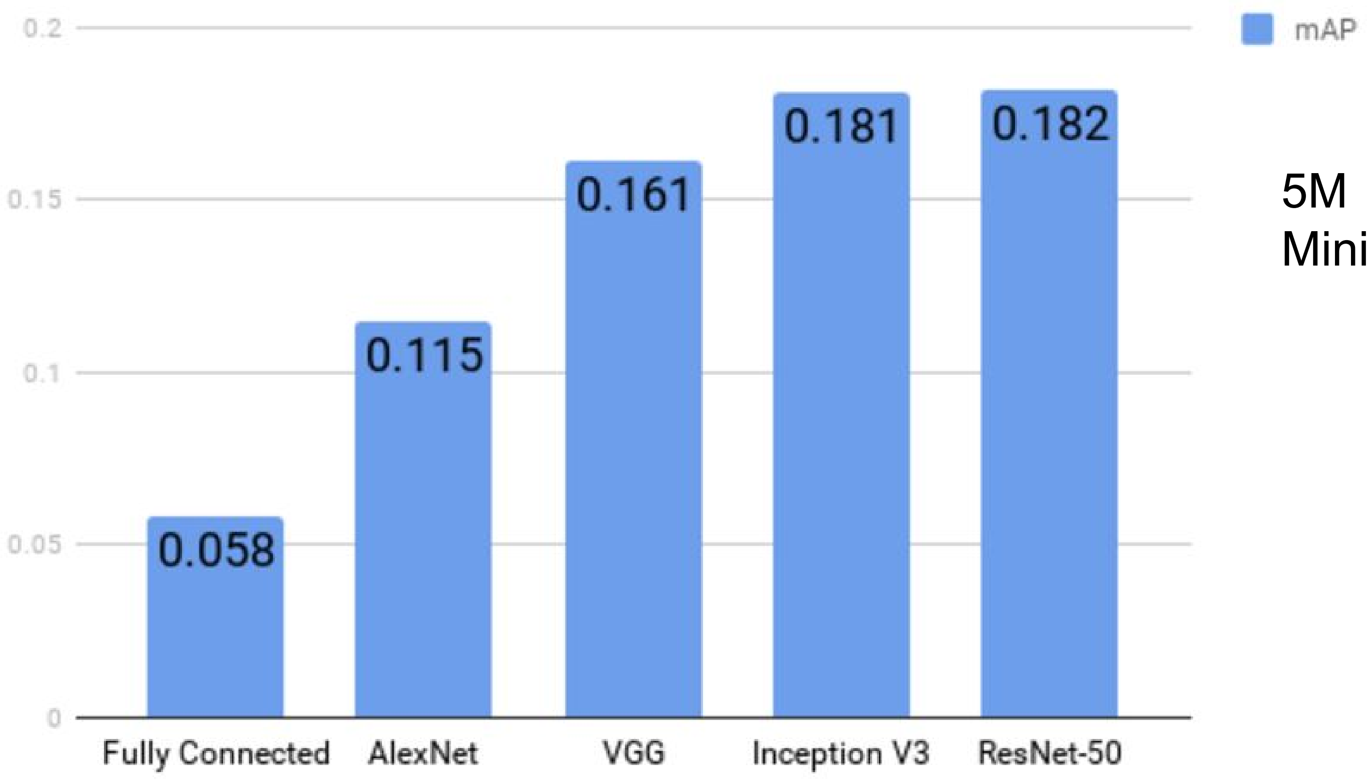


Image from Kaiming He's ResNet presentation



Architectures - Results



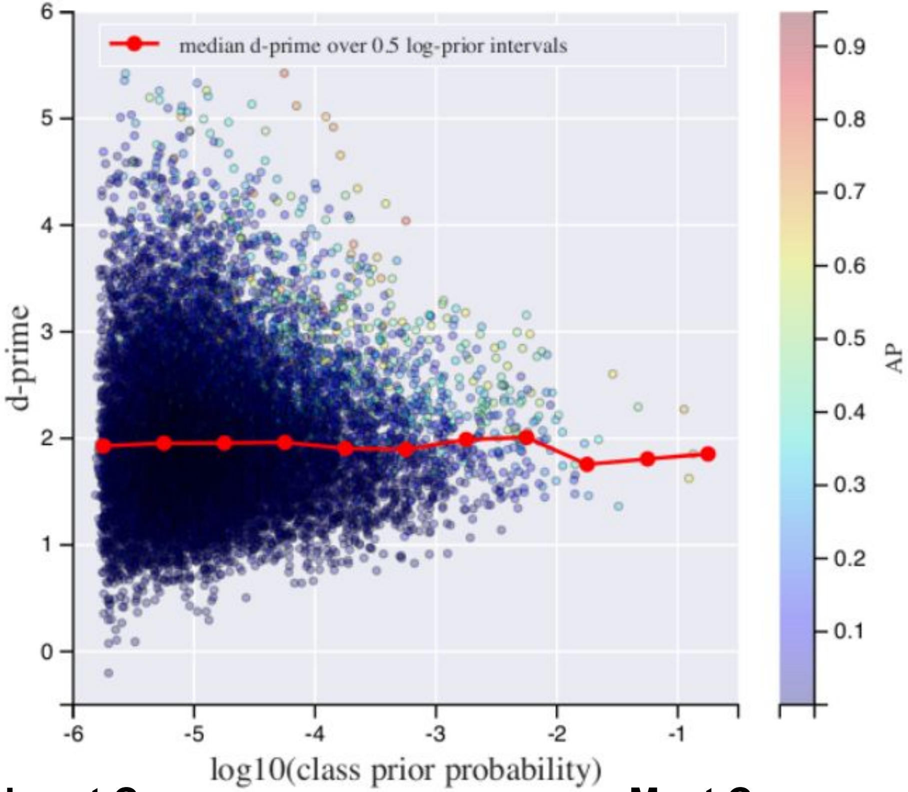
5M Minibatches

Training Size - Results

- Model Used: ResNet-50 (16M mini batches)

DPrime vs Prior

Good



Less Good

Least Common

Most Common

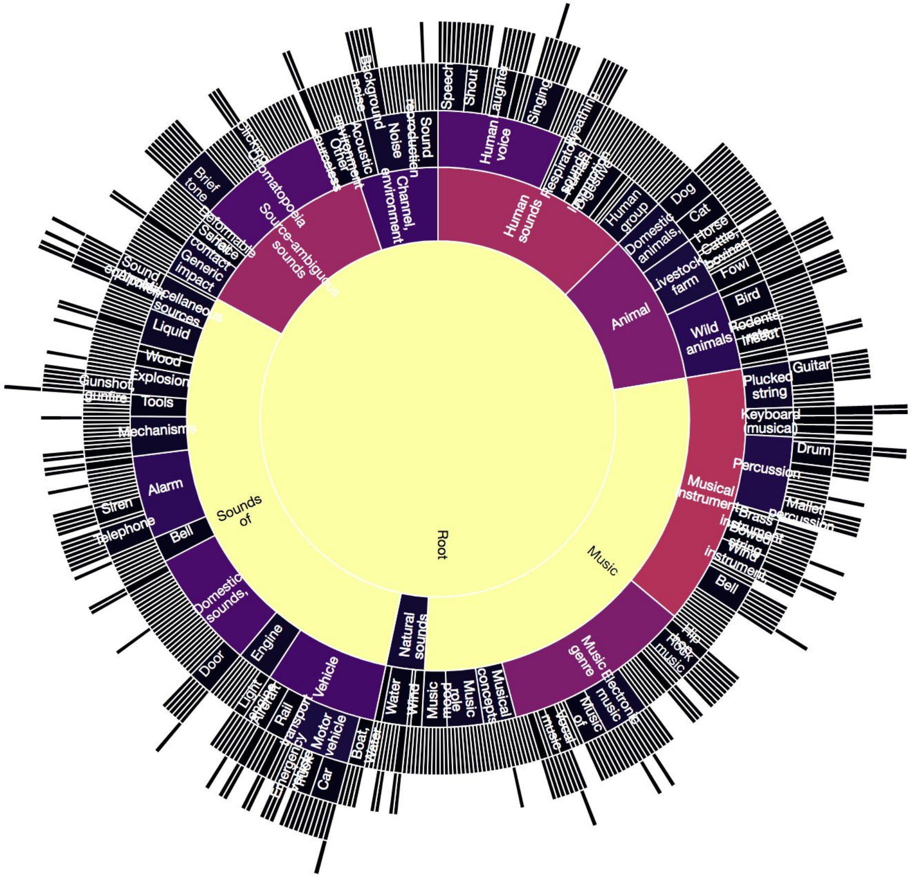
What's Next?

- Web video tags are not sound terms
  - We want the soundtrack described
  - We need a set of sound-description terms

The AudioSet Ontology

github.com/audioset/ontology

- Need a set of sound events
  - 635 “sound” terms in 7 categories
- Start from **Hearst patterns**:
  - *“.. sounds, such as X ..”*
- Refined via:
  - Other sound event lists (Salamon’14, Burger’12,..)
  - Feedback from raters
  - Manual inspection...



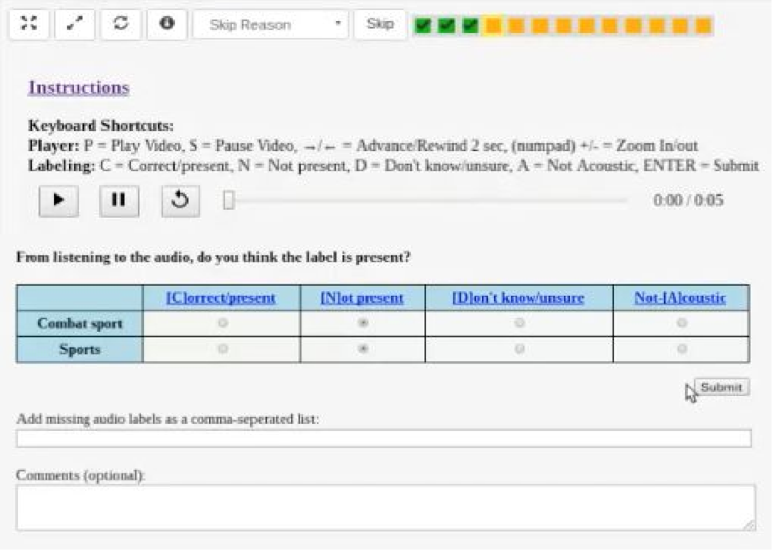
More About The AudioSet Ontology

- Ontology Class Set goals:
  - Not too fine: Non-expert can discriminate consistently
  - Not too few: Cover all normally-encountered sounds
- Evolution
  - Merges: “Tire squeal” + “Skidding”
  - Deletions: “Sidetone”
  - Additions: “Ukulele”, “Stairs”


```
{  "id": "/m/0160x5",
   "name": "Digestive",
   "description": "Sounds associated with the human function of
eating and processing nutrition (food).",
   "citation_uri": "",
   "positive_examples": [],
   "child_ids": ["/m/03cczk", "/m/07pdhp0", "/m/0939n_", "/m/01g90h",
   "restrictions": ["abstract"] },
{  "id": "/m/03cczk",
   "name": "Chewing, mastication",
   "description": "Food being crushed and ground by teeth.",
   "citation_uri": "http://en.wikipedia.org/wiki/Mastication",
   "positive_examples": ["youtu.be/EBnrA85wsc4?start=530&end=540", "y
   "child_ids": [],
   "restrictions": [] },
```

AudioSet Labeled Data

- Verification task
  - Using metadata, identify videos that may contain sound class X
  - Check for other possible classes
  - Extract random 10 s excerpt
  - Ask labeler Present/Absent/Unsure



When Metadata Fails

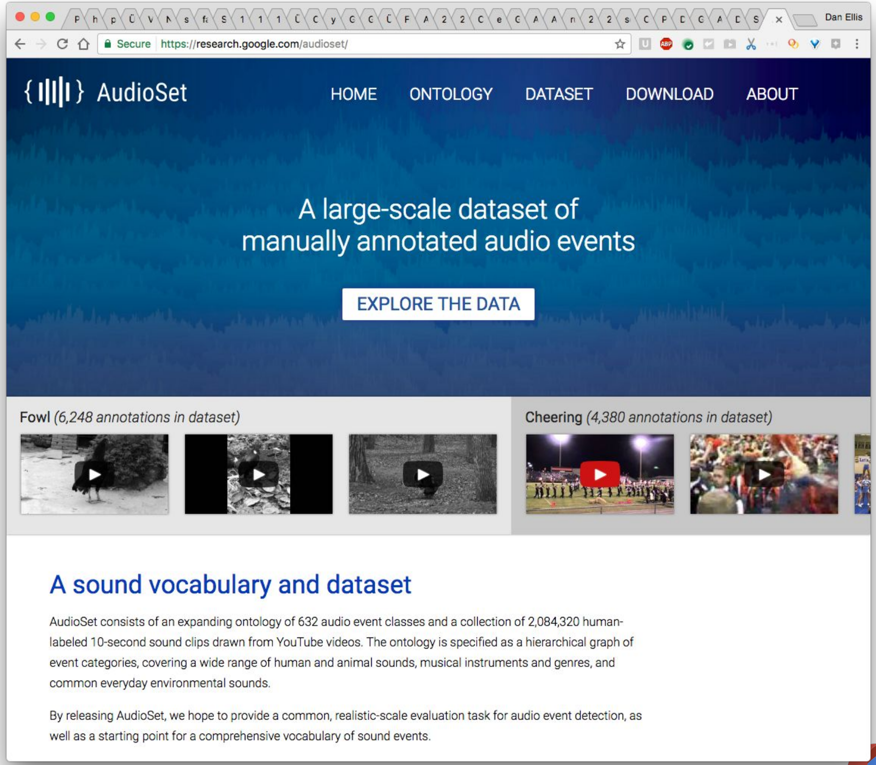
- For obscure sounds, metadata fails to find a large number of good candidates

- Exemplar-based Mining (assumes a few 10s segments for class):
    1. Extract frame-level embeddings using YT-100M bottleneck layer
    2. Cluster the frame-level embeddings to find frames shared across segments
    3. Use those frames in multiquery-by-example search over a millions of YT videos
        - Retrieval score is average distance to individual example frames from step 2
    4. Present retrieved frames (padded to 10s) to labeler for verification

- **Pro:** recovers a more diverse collection of new positives and difficult negatives for rare classes
- **Con**: sampling biased by existing models (still not as diverse as desired)

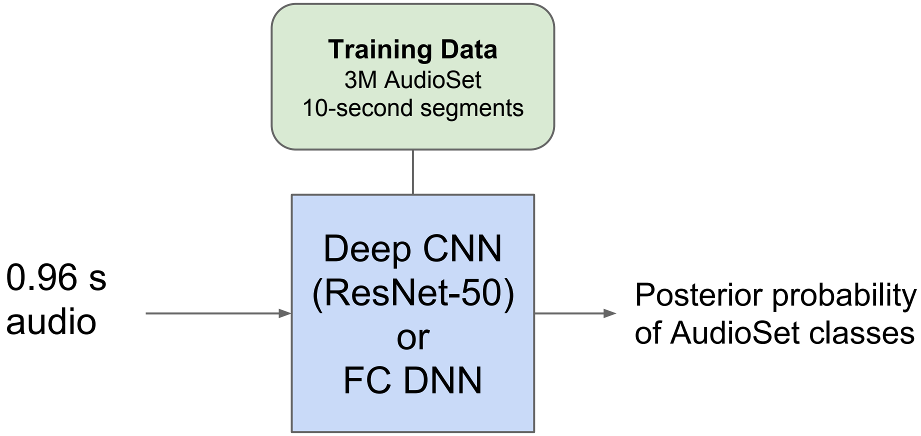
AudioSet Data Release

- A large-scale collection of Labeled sound examples
  - Like ImageNet for sounds
- 2M+ ten-second excerpts from high-viewcount YT videos (1000x smaller than YT-100M But strongly labeled)
- At least 120 examples for 500+ classes



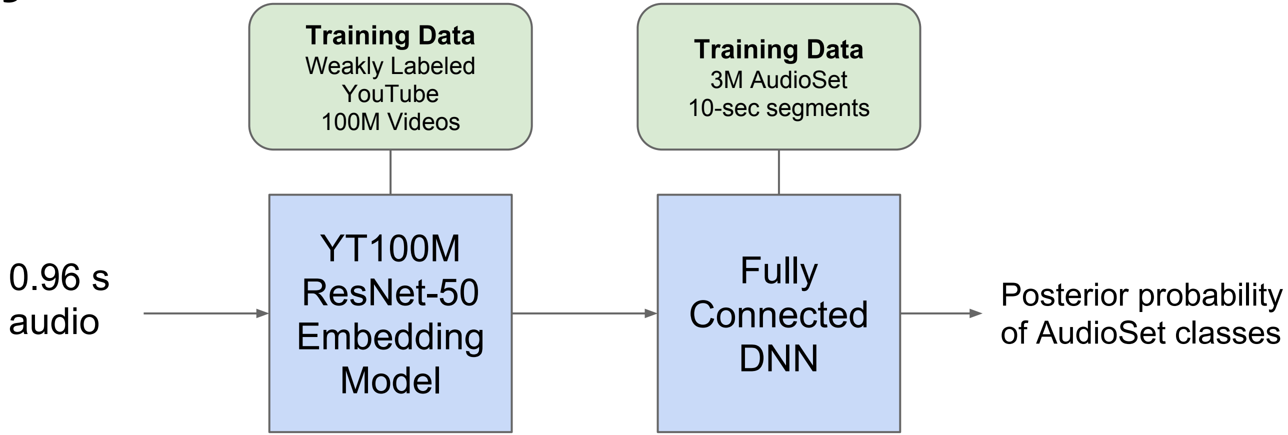
Training Classifiers on AudioSet

Strategy 1



Training Classifiers on AudioSet

Strategy 2



AudioSet Performance

- **Training:** 3M AudioSet 10s segments, 527 classes
- **Evaluation:** explicit hard negatives (average prior: 0.282)

Feature	Model	EER	mAP
Log Mel	Fully connected	33.3%	0.445
Log Mel	ResNet-50	**24.5%**	**0.605**
YT100 Embeddings	Fully connected	25.7%	0.580

Strategy 1

Strategy 2

- Convolutional ResNet model huge improvement over fully connected
- Transfer of embedding from large-scale weakly labeled model does not help overall (but greatly reduces training data requirements in semi-supervised experiments)

Complication #1: Extreme Class Imbalance

- **Problem:** priors range from 0.0001 (e.g. toothbrush, gargling, creak) to 0.5 (e.g. music, speech, vehicle)
    - Results in poor score calibration across classes
    - High-prior classes like speech and music are always strongest detections

- **1990s libsvm solution:** per-class loss weights that balance positive and negative examples
    - Simply does not work at our level of imbalance and model complexity
    - Shared network: one toothbrush example is not worth 5000 speech examples

Complication #1: Extreme Class Imbalance

- **Solution that works:** per-class loss weights that balance to mean prior (~0.003), not 0.5
  - We weight class C *positive* example loss contribution by

Mean prior

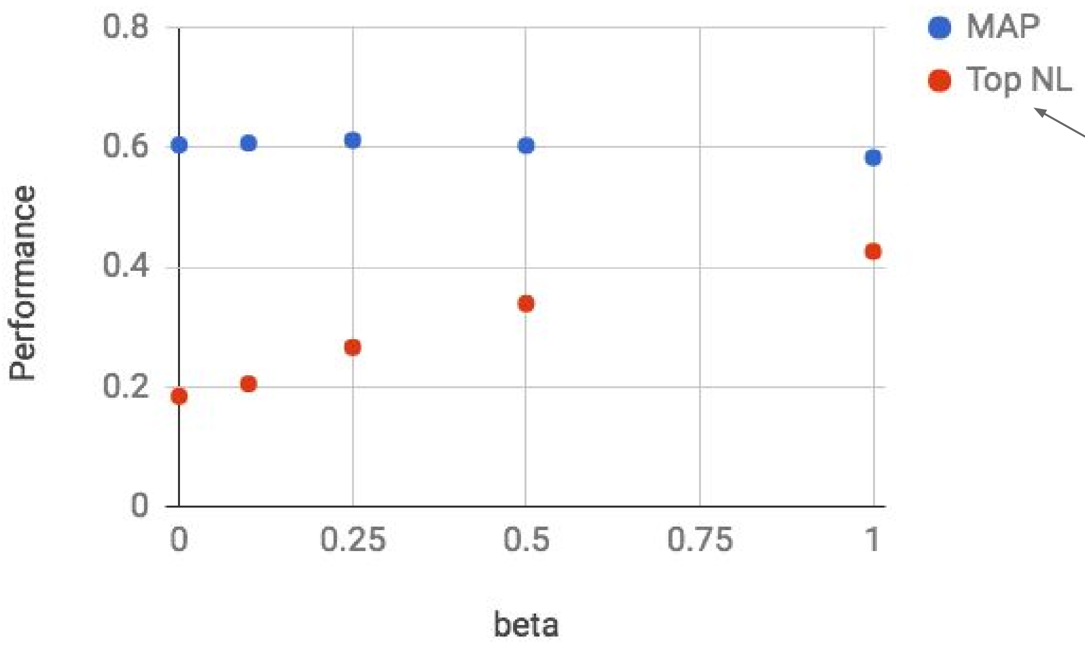
Hyperparameter in [0, 1]

$$\left[\frac{\bar{p}}{p_C} \cdot \frac{1 - p_C}{1 - \bar{p}}\right]^{\beta}$$

Class prior

  - *Negative* examples not weighted.
  - Exponent hyperparameter allows easy backing off from full balancing

Weighted Loss Performance



Average accuracy of the N_L highest scoring classes, where N_L=number of ground truth labels for segment

- Slight improvement to mAP for beta = 0.1, 0.25
- More than doubling of prediction accuracy with full weighting (beta = 1)

Complication #2: Weak Labels

- **Problem:** AudioSet segments are still weakly labeled:
  - Positive label implies event occurs in 10-sec segment, but we do not know extent
- **Solution:** apply simple label refinement of training data:
  1. Train model on original data
  2. For training segment $S = \{x_i\}$ with label L, compute max-normalized frame scores
  $$n_i = \frac{P(L|x_i)}{\max_{x \in S} P(L|x)}$$
  3. Discard frames where n_i falls below some prescribed threshold
- **The Catch:** target class needs to be most prevalent sound in present in labeled segments (relative to prevalence in set as a whole)

Weak Label Refinement Demo

Without refinement
Speech activated throughout

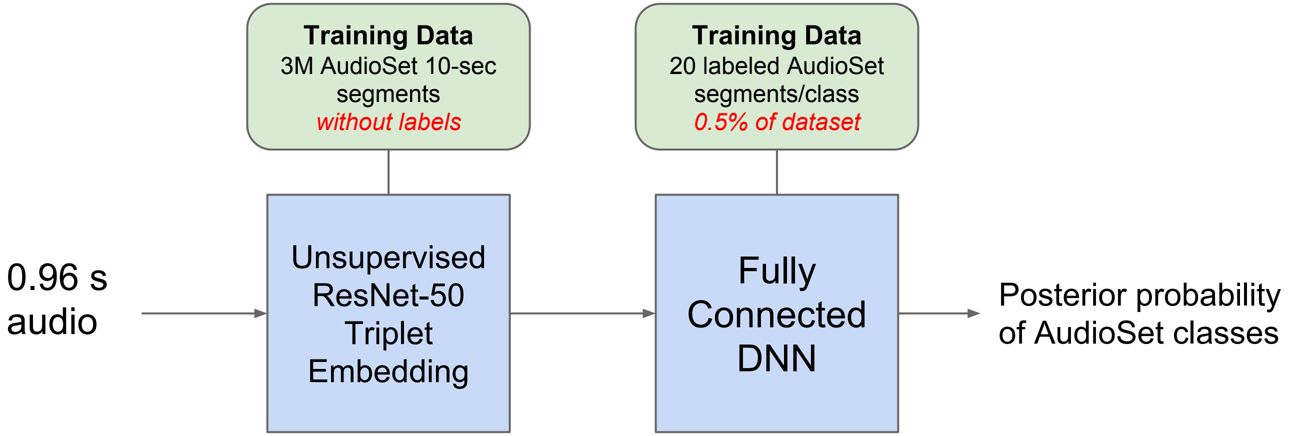
With refinement (0.66 threshold)
Speech activated only when speaking or breathing (difficult confound)

Unsupervised Triplet Loss Embedding

- **The Idea:** train triplet loss embedding models using unsupervised (a.k.a. self-supervised) constraints:
    1. Noise corrupted audio retains the categorical content of the clean signal.
    2. Sound is transparent: mixing two sound classes results in an (often-natural sounding) example of both classes.
    3. Sound classes are translation invariant in time and, to some extent, frequency.
    4. Sounds in close proximity or in same source content are likely to be categorically similar

- When expressed as triplets, trivial to combine all constraints into single huge convolutional network

Semi-Supervised AudioSet Classifier



Semi-Supervised AudioSet Performance

% Data Labeled	Feature	Model	EER	mAP
100%	Log Mel	Fully connected	33.3%	0.445
100%	Log Mel	ResNet-50	24.5%	0.605
0.5%	Log Mel	Fully connected	40.6%	0.338
0.5%	Log Mel	ResNet-50	37.7%	0.385
0.5%	Unsup. Triplet Embedding	Fully connected	34.0%	0.429

AudioSet Demo Video

Future Work

- Transient Events
- Sound Mixtures
- Other Data Sources
  - Closed captions
  - Sound Effects Databases
  - Direct solicitation
  - Other modalities & label sources