Coincidence, Categorization, and Consolidation: Learning to Recognize Sounds with Minimal Supervision

Aren Jansen, Daniel P. W. Ellis, Shawn Hershey, R. Channing Moore, Manoj Plakal, Ashok C. Popat, Rif A. Saurous

Google Research

Getting Started On A New ML Application

- **Goal:** Collect N examples for each of K classes

<u>Case #1: Have Unlabeled Data</u>

- **Common Strategy:** Randomly sample examples for rating
- **Problem:** biased class distribution and abundance of out-of-set classes

<u>Case #2: No Unlabeled Data</u>

- **Common Strategy:** Collect artificially prompted examples
- **Problem:** not fully representative of data in deployment setting

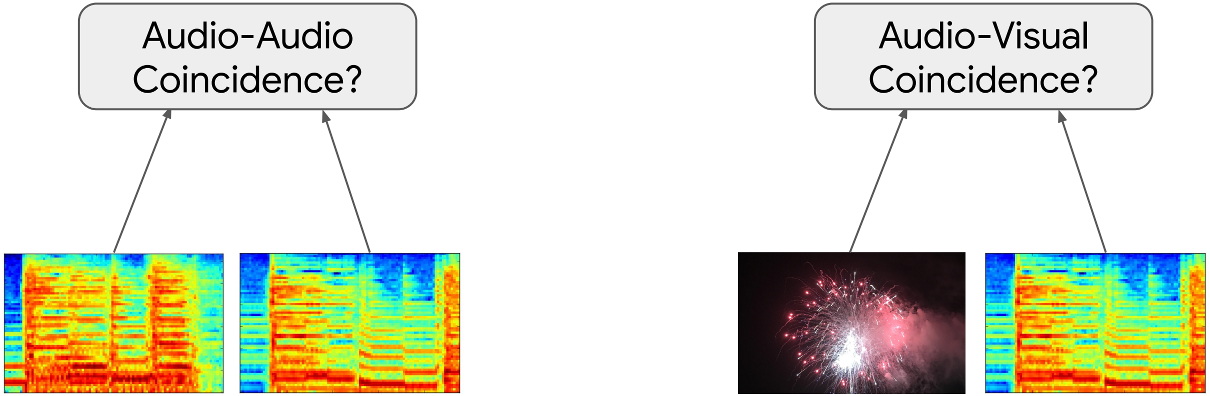
Inspiration from Infant/Child Cognitive Learning

- Humans enter the world with no ability to:
    - Track and recognize objects
    - Recognize speech and environmental sounds
- Abilities only emerge throughout first year after several months of largely unsupervised exposure to natural stimuli
- Once two-way communication is established:
    - Children know what they don't know and ask for labels for novel classes
    - **However:** they don't need a label for every instance

Coincidence, Categorization, and Consolidation

Goal: Go from unlabeled dataset to semantic classifier similar to how children acquire cognitive skills:

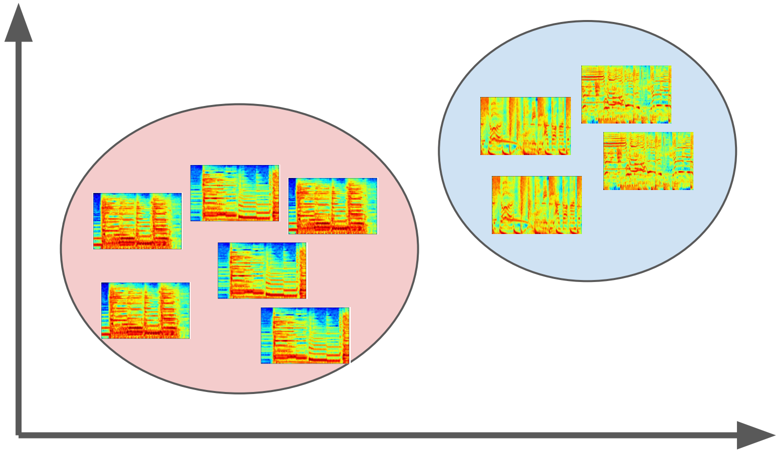
1. **Coincidence:** observe which stimuli do and don't coincide to learn a semantic representation



Coincidence, Categorization, and Consolidation

Goal: Go from unlabeled dataset to semantic classifier similar to how children acquire cognitive skills:

2. **Categorization:** Apply cluster-based category discovery methods to representation and reinforce with clustering loss

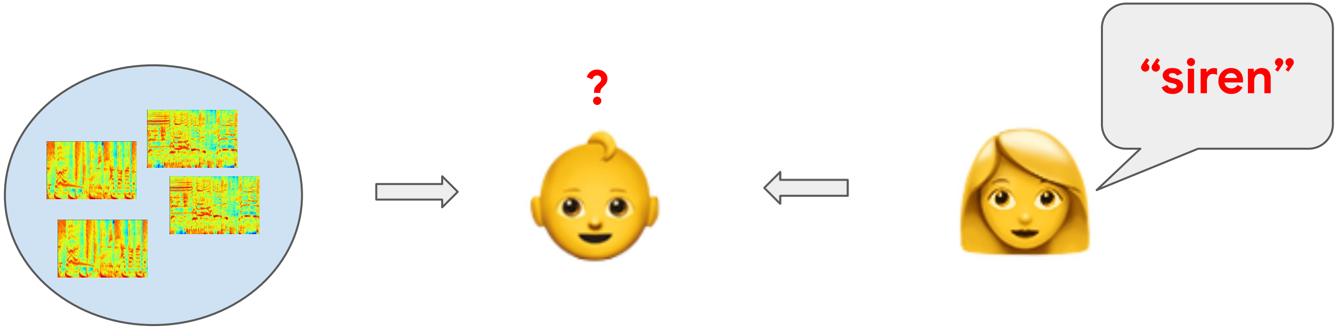


Google Research

Coincidence, Categorization, and Consolidation

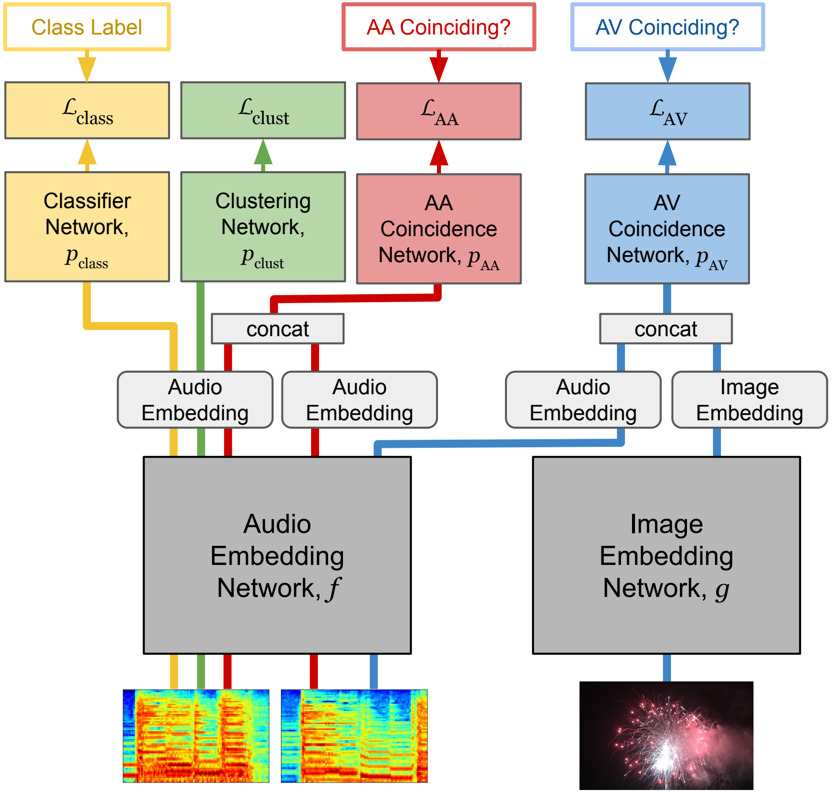
Goal: Go from unlabeled dataset to semantic classifier similar to how children acquire cognitive skills:

3. **Consolidation:** Solicit semantic label for each cluster and train an additional classifier layer.



Plus: Do It With a Single Network

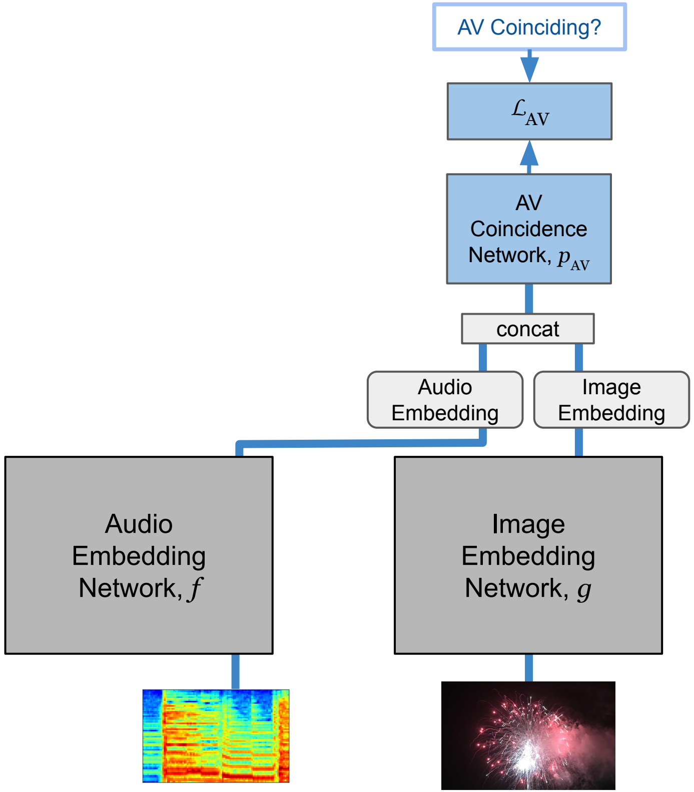
- **Training data:**
  - (audio, audio) pairs
  - (audio, image) pairs
  - either nearby in time or not
- **Result:**
  - Audio and image embeddings
  - Clustering network
  - Semantic classifier



Google Research

Curriculum Stage #1: AV Coincidence Prediction

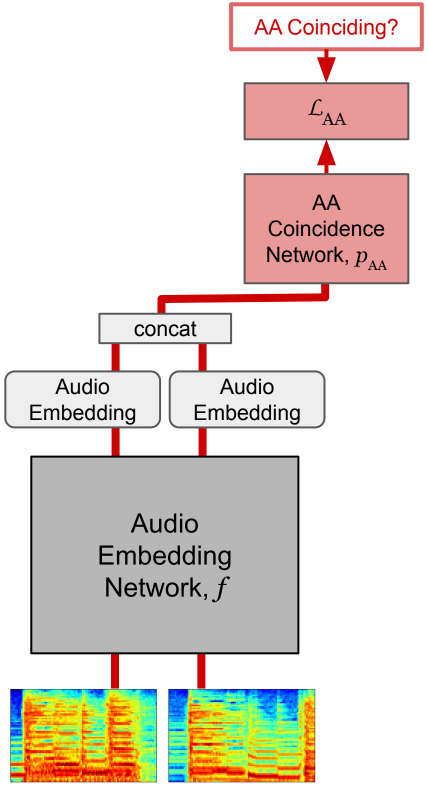
- **Baseline: AV Correspondence**
  - Predict whether AV frames overlap
  - “Look, Listen, and Learn” (2017)
- **We generalize to AV Coincidence**
  - Predict whether AV frames temporally proximal (< ΔT)
- **Why?**
  - Do not need to see source making sound
  - Allows unification with audio-only coincidence prediction
- **Other changes:**
  - VGG → ResNet-50
  - Random negatives → all-pairs batch construction



Google Research

Curriculum Stage #2: AA + AV Coincidence Prediction

- Like AV Coincidence prediction, but with two audio inputs and dedicated prediction network
- Conceptually equivalent to our temporal proximity triplet embedding technique from [Jansen et al., ICASSP 2018]



Curriculum Stage #3: AV + AA + Entropy-Based Clustering

- Entropy-based loss function and optimization with SGD:

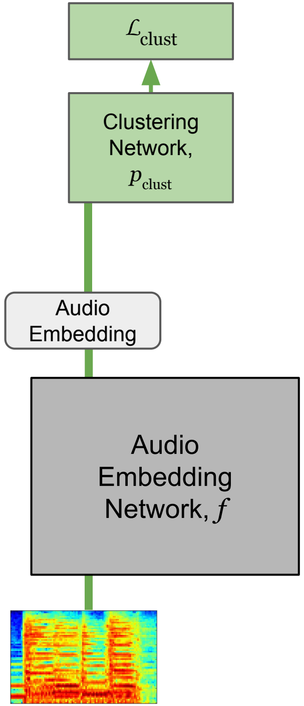
Confident Assignments

$$\mathcal{L}_{\text{clust}}(X) = \frac{1}{B}\sum_{i=1}^{B} H[p_{\text{clust}}(f(x_i))]$$

$$-\gamma H\left[\frac{1}{B}\sum_{i=1}^{B} p_{\text{clust}}(f(x_i))\right]$$

Diverse Assignments

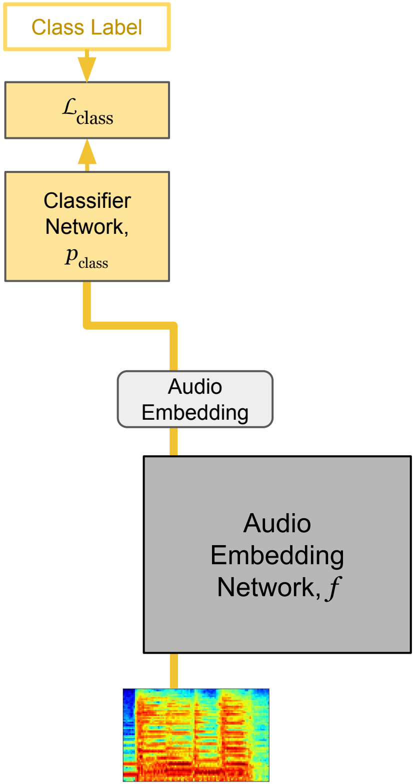
- Easily scales to 1M clusters and all in TensorFlow
- Out-of-sample extension is just regular forward pass



Google Research

Curriculum Stage #4: Weakly-Supervised Classification

- Solicit label for one random example per cluster
- Propagate label to unlabeled examples in each cluster
- Add classifier network to audio embedding
- Apply standard cross-entropy classification loss using weak labels

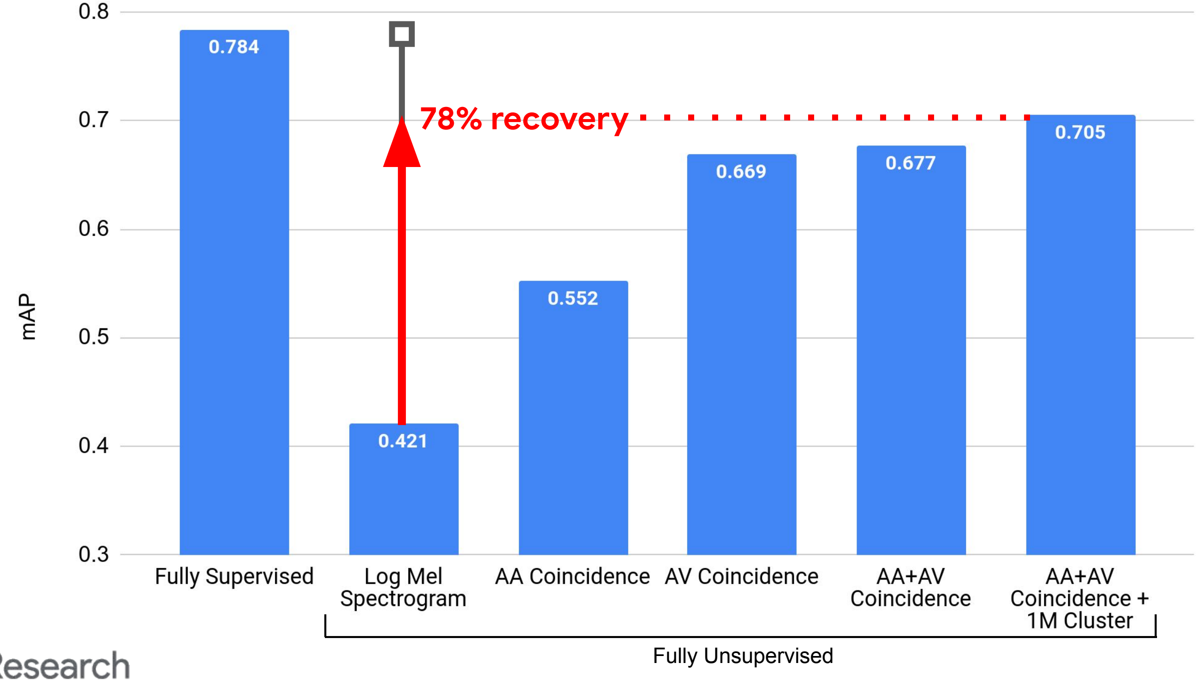


AudioSet Benchmark (g.co/audioset)

- **AudioSet:** 2M YouTube training segments, 527 classes, prior imbalance up to 10,000:1
- **Embedding Models:** ResNet-50 ⟶ 128-dimensional embedding
- **Topline Representation:** fully-supervised semantic embedding (trained with triplet loss)
- **Baseline Representation:** input log mel spectrogram features

Eval #1: Query-By-Example

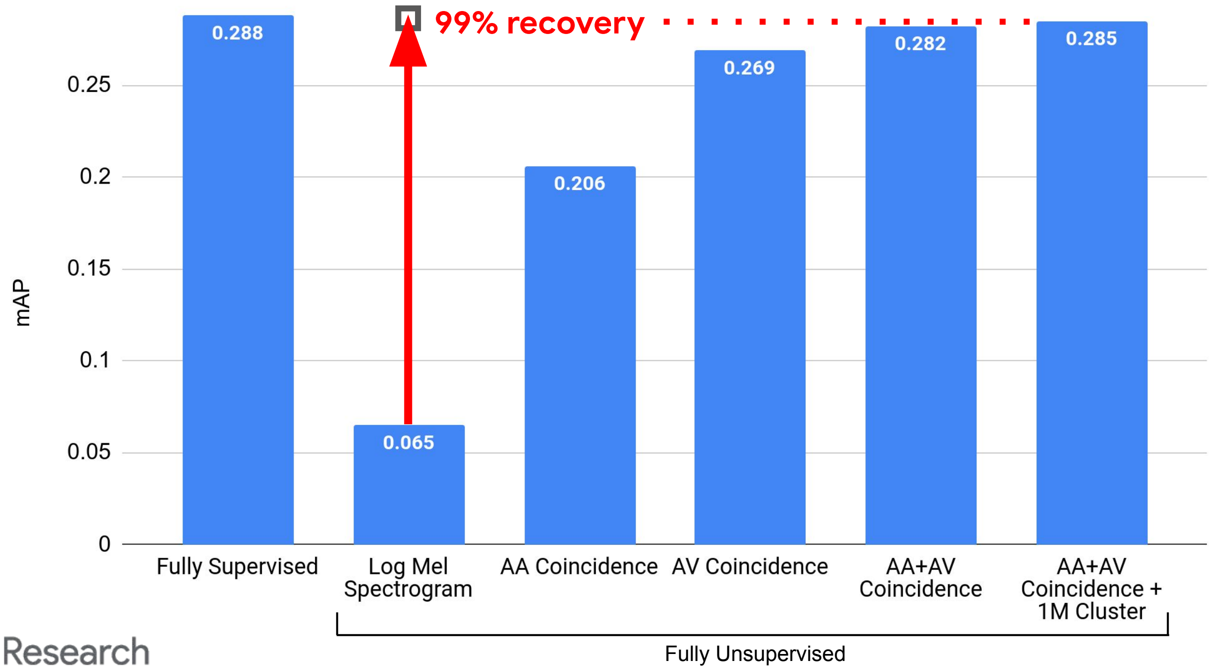
- **Eval:** Rank same/different class example pairs by cosine distance
- **Measures:** Intrinsic semantic quality of representation



Unsupervised representation recovers 78% of the fully supervised gap!

Eval #2: Shallow Classifier

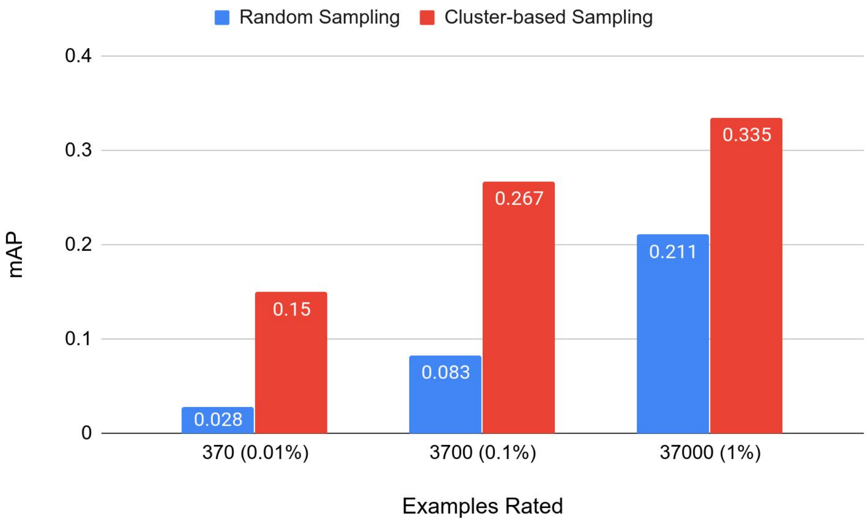
- **Eval:** Shallow fully-connected (FC1x512) classifier holding representation fixed
- **Measures:** Representation support of downstream classification tasks



Unsupervised representation recovers 99% of the fully-supervised gap!

Eval #3: Unsupervised Active Learning

1. Cluster dataset using unsupervised semantic representation
2. Label N biggest clusters by rating a random example from each
3. Train classifier with noisy cluster-based labels



Unsupervised active learning reduces label requirement by more than 10X

Conclusions

- In-domain unsupervised audio embedding reaches supervised performance
- Unsupervised active learning gives 10X reduction in label requirements
- Lessons for audio ML and beyond:
  - Collect unlabeled data when it is free/cheap
  - Collect second modality when you can
  - Cluster-based sampling > random sampling (given good representation)