Transfer Learning From Youtube Soundtracks to Tag Arctic Ecoacoustic Recordings

Enis Berk Çoban, Dara Pir, Richard So, Michael I Mandel

ICASSP 2020
Problems

- Arctic-Boreal forests are warming at twice the global average
- Bird migrations and reproductive success are impacted
Problems

- Oil and Gas Extraction
- Frequency of vehicle usage is increasing
Solution

- Ecoacoustic monitoring
- Machine learning for data processing
- Transfer Learning
Data and Experiments
Outline

Data:
- 3 months of nature sounds from Alaska
- 8 categories for labeling

Experiments:
1. Out of box usage of an Audio classifier (0.7 AUC)
2. Training Classifiers with the Audio classifier results (0.77 AUC)
3. Training Classifiers with lower level embeddings (0.86 AUC)

Visualization:
- Songbird Predictions for 7 locations
Data

- Collected by Taylor Stinchcomb*
- The Colville River in Alaska
- 3 Months (June, July, and August of 2016)

<table>
<thead>
<tr>
<th>Tag</th>
<th>Sample Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wind</td>
<td>641</td>
</tr>
<tr>
<td>Cable Noise</td>
<td>456</td>
</tr>
<tr>
<td>Songbird</td>
<td>409</td>
</tr>
<tr>
<td>Running Water</td>
<td>210</td>
</tr>
<tr>
<td>Water Bird</td>
<td>196</td>
</tr>
<tr>
<td>Insect</td>
<td>190</td>
</tr>
<tr>
<td>Rain</td>
<td>102</td>
</tr>
<tr>
<td>Aircraft</td>
<td>28</td>
</tr>
</tbody>
</table>

Songbird
Aircraft
Insect
- **Exp 1**: VGGish + Audio Set
- **Exp 2**: VGGish + Audio Set + Traditional Classifiers
- **Exp 3**: VGGish + Traditional Classifiers
VGGish

- Deep convolutional network adapted to audio from VGG object recognizer
- Trained on the Youtube-100M dataset
- 128-dimensional embedding vectors

Audio Set Classifier

- Trained on hierarchically organized sound events from Youtube-100M*
- Attention-based

- Log mel spectrograms of 960 ms sound excerpts
  - Size 96 × 64

---

• Combining multiple Audio Set labels into each event category
Exp 1 - Manual Audio Set Mapping

**Songbird:** Bird; Owl; Bird vocalization, bird call, ...

**WaterBird:** Duck; Goose; Quack; Frog; ...

**Insect:** Fly, housefly; Insect; Bee, wasp, etc.; ...

**Aircraft:** Engine; Fixed-wing aircraft, airplane; ...

**Running Water:** Waterfall; Waves, surf

**Cable:** Bang; Slap, smack; Whack, thwack; ...

**Wind:** Wind; Howl

**Rain:** Rain; Raindrop; Rainonsurface
Exp 1 - Manual Audio Set Mapping

**Songbird**: Bird; Owl; Bird vocalization, bird call, bird song; Pigeon, dove; Coo; Chirp, tweet; Squawk; Bird flight, flapping wings; Gull, seagull; Chirp tone; Hoot

**WaterBird**: Duck; Goose; Quack; Frog; Croak; Caw

**Insect**: Fly, housefly; Insect; Bee, wasp, etc.; Buzz; Mosquito; Cricket; Rustle

**Aircraft**: Engine; Fixed-wing aircraft, airplane; Aircraft engine, Propeller, airscrew; Aircraft; Helicopter

**Running Water**: Waterfall; Waves, surf

**Cable**: Bang; Slap, smack; Whack, thwack; Smash, crash; Breaking; Knock; Tap; Thump, thud; Whip; Flap; Clip-clop

**Wind**: Wind; Howl

**Rain**: Rain; Raindrop; Rain on surface
Exp 1 - Manual Audio Set Mapping

<table>
<thead>
<tr>
<th>Tag</th>
<th>NPos</th>
<th>Bulbul</th>
<th>Manual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wind</td>
<td>641</td>
<td>0.70</td>
<td>0.66</td>
</tr>
<tr>
<td>Cable noise</td>
<td>456</td>
<td>0.70</td>
<td>0.65</td>
</tr>
<tr>
<td>Songbird</td>
<td>409</td>
<td>0.86</td>
<td>0.70</td>
</tr>
<tr>
<td>Running water</td>
<td>210</td>
<td>0.70</td>
<td>0.57</td>
</tr>
<tr>
<td>Water bird</td>
<td>196</td>
<td>0.65</td>
<td>0.59</td>
</tr>
<tr>
<td>Insect</td>
<td>190</td>
<td>0.58</td>
<td>0.66</td>
</tr>
<tr>
<td>Rain</td>
<td>102</td>
<td>0.56</td>
<td>0.44</td>
</tr>
<tr>
<td>Aircraft</td>
<td>28</td>
<td>0.66</td>
<td>0.52</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Audio Set</th>
<th>VGGish10</th>
<th>VGGish1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.85 (gp)</td>
<td>0.90 (gp)</td>
<td>0.91 (nn)</td>
</tr>
<tr>
<td>0.80 (rbf)</td>
<td>0.87 (gp)</td>
<td>0.86 (gp)</td>
</tr>
<tr>
<td>0.77 (gp)</td>
<td>0.83 (nn)</td>
<td>0.86 (nn)</td>
</tr>
<tr>
<td>0.85 (gp)</td>
<td>0.92 (nn)</td>
<td>0.89 (nn)</td>
</tr>
<tr>
<td>0.74 (gp)</td>
<td>0.76 (nn)</td>
<td>0.77 (rbf)</td>
</tr>
<tr>
<td>0.79 (nn)</td>
<td>0.87 (lsvm)</td>
<td>0.82 (lsvm)</td>
</tr>
<tr>
<td>0.81 (rbf)</td>
<td>0.85 (gp)</td>
<td>0.82 (gp)</td>
</tr>
<tr>
<td>0.78 (nn)</td>
<td>0.86 (ab)</td>
<td>0.52 (gp)</td>
</tr>
</tbody>
</table>

- **VGGish + Audio Set**
- **Bulbul**: attention-based state-of-the-art dedicated bird detector

The AUC score for each category given predictions of a single Audio Set label

Unrelated labels predict certain categories successfully
Exp 2 - Traditional Classifiers on Top of Audio Set Labels

- Log-mel Spectrogram
- Audio Set Classifier
- 527 Audio Set Classes
- Traditional Classifiers
- 8 Event Categories

Exp 1

Exp 2

Exp 3
Exp 2 - Traditional Classifiers on Top of Audio Set Labels

<table>
<thead>
<tr>
<th>Tag</th>
<th>NPos</th>
<th>Bulbul</th>
<th>Manual</th>
<th>Audio Set</th>
<th>VGGish10</th>
<th>VGGish11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wind</td>
<td>641</td>
<td>0.70</td>
<td>0.66</td>
<td>0.85 (gp)</td>
<td>0.90 (gp)</td>
<td>0.91 (nn)</td>
</tr>
<tr>
<td>Cable noise</td>
<td>456</td>
<td>0.70</td>
<td>0.65</td>
<td>0.80 (rbf)</td>
<td>0.87 (gp)</td>
<td>0.86 (gp)</td>
</tr>
<tr>
<td>Songbird</td>
<td>409</td>
<td>0.86</td>
<td>0.70</td>
<td>0.77 (gp)</td>
<td>0.83 (nn)</td>
<td>0.86 (nn)</td>
</tr>
<tr>
<td>Running water</td>
<td>210</td>
<td>0.70</td>
<td>0.57</td>
<td>0.85 (gp)</td>
<td>0.92 (nn)</td>
<td>0.89 (nn)</td>
</tr>
<tr>
<td>Water bird</td>
<td>196</td>
<td>0.65</td>
<td>0.59</td>
<td>0.74 (gp)</td>
<td>0.76 (nn)</td>
<td>0.77 (rbf)</td>
</tr>
<tr>
<td>Insect</td>
<td>190</td>
<td>0.58</td>
<td>0.66</td>
<td>0.79 (nn)</td>
<td>0.87 (lsvm)</td>
<td>0.82 (lsvm)</td>
</tr>
<tr>
<td>Rain</td>
<td>102</td>
<td>0.56</td>
<td>0.44</td>
<td>0.81 (rbf)</td>
<td>0.85 (gp)</td>
<td>0.82 (gp)</td>
</tr>
<tr>
<td>Aircraft</td>
<td>28</td>
<td>0.66</td>
<td>0.52</td>
<td>0.78 (nn)</td>
<td>0.86 (ab)</td>
<td>0.52 (gp)</td>
</tr>
</tbody>
</table>

- Test set results using classifiers with best validation set performance
Exp 3 - Traditional Classifiers on VGGish Embeddings

VGGish

Log-mel Spectrogram

Audio Set Classifier

527 Audio Set Classes

8 Event Categories

Traditional Classifiers

526

Exp 1

Exp 2

Exp 3

0 1 7
Exp 3 - Traditional Classifiers on VGGish Embeddings

VGGish-1: Combining embeddings

VGGish

Average or Concatenate

Traditional Classifiers

Exp 3

0
1

8 Event Categories

VGGish-10: Weakly supervised

VGGish

Traditional Classifiers

Max

Exp 3

0
1

7
Exp 3 - Traditional Classifiers on VGGish Embeddings

<table>
<thead>
<tr>
<th>Tag</th>
<th>NPos</th>
<th>Bulbul</th>
<th>Manual</th>
<th>Audio Set</th>
<th>VGGish10</th>
<th>VGGish1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wind</td>
<td>641</td>
<td>0.70</td>
<td>0.66</td>
<td>0.85 (gp)</td>
<td>0.90 (gp)</td>
<td>0.91 (nn)</td>
</tr>
<tr>
<td>Cable noise</td>
<td>456</td>
<td>0.70</td>
<td>0.65</td>
<td>0.80 (rbf)</td>
<td>0.87 (gp)</td>
<td>0.86 (gp)</td>
</tr>
<tr>
<td>Songbird</td>
<td>409</td>
<td><strong>0.86</strong></td>
<td>0.70</td>
<td>0.77 (gp)</td>
<td>0.83 (nn)</td>
<td>0.86 (nn)</td>
</tr>
<tr>
<td>Running water</td>
<td>210</td>
<td>0.70</td>
<td>0.57</td>
<td>0.85 (gp)</td>
<td>0.92 (nn)</td>
<td>0.89 (nn)</td>
</tr>
<tr>
<td>Water bird</td>
<td>196</td>
<td>0.65</td>
<td>0.59</td>
<td>0.74 (gp)</td>
<td>0.76 (nn)</td>
<td>0.77 (rbf)</td>
</tr>
<tr>
<td>Insect</td>
<td>190</td>
<td>0.58</td>
<td>0.66</td>
<td>0.79 (nn)</td>
<td><strong>0.87</strong> (lsvm)</td>
<td>0.82 (lsvm)</td>
</tr>
<tr>
<td>Rain</td>
<td>102</td>
<td>0.56</td>
<td>0.44</td>
<td>0.81 (rbf)</td>
<td><strong>0.85</strong> (gp)</td>
<td>0.82 (gp)</td>
</tr>
<tr>
<td>Aircraft</td>
<td>28</td>
<td>0.66</td>
<td>0.52</td>
<td>0.78 (nn)</td>
<td><strong>0.86</strong> (ab)</td>
<td>0.52 (gp)</td>
</tr>
</tbody>
</table>

- These models perform well enough that we can use them with a certain confidence
- VGGish1 averaging version is reported
Songbird Predictions Over 7 Sites

- “Songbird” Neural Network model trained on VGGish raw embeddings
- Top: June, Middle: July, Bottom: August.
Songbird Predictions Over 7 Sites

USGS
Ryklak 1
Ryklak 2
Colville 2
Ocean Pt
Colville 4
Colville 5
Songbird Predictions Over 7 Sites

- "Songbird" Neural Network model trained on VGGish raw embeddings
- Top: June
- Middle: July
- Bottom: August
Conclusions

● **Best technique** - Classical ML models with VGGish embeddings as input

● **Results** - AUC above 80% for all categories except one

● **Exception** - Water birds (we grouped waterfowl together with shorebirds)

● **General** - This general model performs on par with Bulbul, which is specialized for songbird, but much better on the other tags
Future Work

- Break categories down into a finer granularity, species level
- Identify important events in phenology of bird communities
- Measure human-generated noise affecting caribou herds
Acknowledgements

This material is based upon work supported by the National Science Foundation (NSF) grant OPP-1839185. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.
Thank you