Goal: Self-supervised learning

Pretext Task:

Hold out one part of an image or video and predict it from the rest.
Self-supervised learning with sound

Analogously, hold out the sound from video and predict it from the images.
Audio is invariant to many visual transformations
Audio is invariant to many visual transformations

Image space

Audio space

![Image of ocean scene with moon and camera icon]
Audio is invariant to many visual transformations
Audio is invariant to many visual transformations

Image space

Audio space
Audio is invariant to many visual transformations
Experimental Design

Common sources

Image

Audio
Representing ambient sound

- Predict sound from single image
- Use sound textures [McDermott]
Representing ambient sound

\[ \frac{1}{N} \sum_{i} x_i \]

Mean power

Frequency channel
Representing ambient sound

\[
\frac{1}{N} \| x \ast f_j \|^2
\]

[McDermott & Simoncelli]

Sound texture

Moments
Representing ambient sound

Correlation between frequency channels

Sound texture

[McDermott & Simoncelli]
Representing ambient sound

- Sound texture (McDermott & Simoncelli)
- Moments
- Filter responses
- Correlations
Training Data

- Flickr video dataset.
- 180K videos, 10 random frames from each.
- Trained from scratch

Video frame

ConvNet
Predicting sound (clustering)

Video frame

ConvNet

Sound feature vectors
Predicting sound

Audio cluster

ConvNet

Video frame

K-means or LSH
What information is captured by this sound representation?

Top audio clips for one cluster (of 30).
Top audio clips for one cluster (of 30).
What did the network learn?
What did the network learn?

PASCAL VOC 2007

“person”
PASCAL VOC Classification

- Sound [Doersch15]: 46.7%
- Context [Wang15]: 46.1%
- Tracking [Agrawal15]: 42.2%
- Egomotion [Agrawal15]: 31.2%
- Spectrum only: 44.0%
- Visual clusters: 37.5%
- ImageNet [Krizhevsky12]: 65.5%
### SUN397 Scene Recognition

<table>
<thead>
<tr>
<th>Component</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sound</td>
<td>22.5%</td>
</tr>
<tr>
<td>Sound</td>
<td>22.2%</td>
</tr>
<tr>
<td>Tracking</td>
<td>18.7%</td>
</tr>
<tr>
<td>Egomotion</td>
<td>11.3%</td>
</tr>
<tr>
<td>Places</td>
<td>42.1%</td>
</tr>
</tbody>
</table>

*References:*
- [Doersch15] [Wang15] [Agrawal15] [Zhou14]*
<table>
<thead>
<tr>
<th>Method</th>
<th>VOC Cls. (%mAP)</th>
<th>SUN397 (%acc.)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>max5 pool5 fc6 fc7</td>
<td>max5 pool5 fc6 fc7</td>
</tr>
<tr>
<td>Sound (cluster)</td>
<td>36.7 45.8 44.8 44.3</td>
<td>17.3 22.9 20.7 14.9</td>
</tr>
<tr>
<td>Sound (binary)</td>
<td>39.4 46.7 47.1 47.4</td>
<td>17.1 22.5 21.3 21.4</td>
</tr>
<tr>
<td>Sound (spectrum only)</td>
<td>35.8 44.0 44.4 44.4</td>
<td>14.6 19.5 18.6 17.7</td>
</tr>
<tr>
<td>Texton-CNN</td>
<td>28.9 37.5 35.3 32.5</td>
<td>10.7 15.2 11.4 7.6</td>
</tr>
<tr>
<td>K-means (Krähenbühl et al.)</td>
<td>27.5 34.8 33.9 32.1</td>
<td>11.6 14.9 12.8 12.4</td>
</tr>
<tr>
<td>Tracking (Wang and Gupta)</td>
<td>33.5 42.2 42.4 40.2</td>
<td>14.1 18.7 16.2 15.1</td>
</tr>
<tr>
<td>Patch pos. (Doersch et al.)</td>
<td>26.8 46.1 - -</td>
<td>9.8 22.2 - -</td>
</tr>
<tr>
<td>Egomotion (Agrawal et al.)</td>
<td>22.7 31.1 - -</td>
<td>9.1 11.3 - -</td>
</tr>
<tr>
<td>ImageNet (Krizhevsky et al.)</td>
<td><strong>63.6 65.6 69.6 73.6</strong></td>
<td>29.8 34.0 37.8 37.8</td>
</tr>
<tr>
<td>Places (Zhou et al.)</td>
<td>59.0 63.2 65.3 66.2</td>
<td><strong>39.4 42.1 46.1 48.8</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>(%mAP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random init. (Krähenbühl et al.)</td>
<td>41.3</td>
</tr>
<tr>
<td>Sound (cluster)</td>
<td>44.1</td>
</tr>
<tr>
<td>Sound (binary)</td>
<td>43.3</td>
</tr>
<tr>
<td>Tracking (Wang and Gupta)</td>
<td>44.0</td>
</tr>
<tr>
<td>Egomotion (Agrawal et al.)</td>
<td>41.8</td>
</tr>
<tr>
<td>Patch pos. (Doersch et al.)</td>
<td>46.6</td>
</tr>
<tr>
<td>Calib. + Patch (Krähenbühl et al.)</td>
<td><strong>51.1</strong></td>
</tr>
<tr>
<td>ImageNet (Krizhevsky et al.)</td>
<td><strong>57.1</strong></td>
</tr>
<tr>
<td>Places (Zhou et al.)</td>
<td>52.8</td>
</tr>
</tbody>
</table>

For 30-way classification ask: 15.8% classification accuracy (vs. 6.6% for choosing the most common label)
Did the model recognize sound sources?

Object-selective units emerge when recognizing scenes [Zhou et al. 2015]
Unit visualizations

Audio label

256 filters

conv5

13
Strengths

• Model learns a representation that conveys semantics.
• Orthogonal to image-based approaches.
• Invariance to visual transformation.
Weaknesses

• Requires characteristics sounds, but not all objects emit sounds.

• Sound can be misleading.
Other Possible Extensions?

- Leveraging data from social media.
  (i.e. using comments on photos/videos as a supervisory signal)