Look, Listen and Learn

Relja Arandjelović, Andrew Zisserman

February 20th, 2018

Greg Rehm, Shahbaz Rezaei
Supervised Learning

Duck
Duck
Not Duck
Not Duck

Supervised Learning

Predictive Model
Self-Supervised Learning

- Uses unlabeled data for training
- Do not need annotators to perform extensive ground truth labeling
- Saves time and $$
- How do you perform it in practice for CNN’s though?
Siamese Networks

The **Distance Function** decides if the output vectors are close enough to be similar.

The **Neural Network** transforms the input into a properties vector.

**Input Data** (image, text, features...)

**Diagram:**
- Two neurons labeled as $X_2$ and $X_1$ connected to two separate ends of a neuron labeled as $NN$ with the question, "Similar/Not?"
- The two $NN$ neurons are connected with a dashed line indicating they are the same network.
Can Learn Temporal Ordering

Image Location Prediction

Unlabeled training image

Train Deep Net to recover relative position

Can a Self-Supervised Machine Learn what a Boy Band is?
Can a Self-Supervised Machine Learn what a Boy Band is?
Machine Might Learn This:

- >4 young males in row
- Silly hair
- Dumb look on face
False Positives Abound
Solution: Add Audio
The Main Idea

Audio-visual correspondence detector network

Vision subnetwork

Fusion layers

Audio subnetwork

Correspond? Yes / No
Convolution Block (From VGG)
Fusion + Classification Layers

softmax
2

fc2 128x2
2

fc1 1024x128
128

concat
1024
Inputs
Want to train with overlapping audio and images
And train with non-overlapping audio and images
Image input

224x224x3
Audio input

257x199x1

log-spectrogram

1 second 48kHz audio
How to compute a spectrogram

.01 sec. window

FFT
How to compute a spectrogram

Half window overlap → FFT
Training Details

- Uses Adam Optimizer
- Trained with 16 GPUs
- Effective batch size is 256 elements
- Augmentation for images with random cropping/flipping/brightness/saturation changes
- Augmentation for audio with random volume changes of 10%
- Sees 60 million frame-audio pairs
Datasets

- **Flickr-SoundNet**
  - Unlabelled dataset of Flickr video
  - Contains over 2 million videos
  - A random subset of 500K was used

- **Kinetics-Sounds**
  - Manually labelled dataset containing 34 classes
    - Playing instruments (guitar, violin, etc.)
    - Using tools (lawn mowing, shoveling snow, etc.)
    - Performing actions (tap dancing, bowling)
  - Clean, but contains noises:
    - Load background music
    - Human voice
    - Videos with soundtrack
3.2 Audio Visual Correspondence

**L3-Net:** Train the whole network in unsupervised manner.

Train the entire network
3.2 Audio-Visual Correspondence (AVC)

**Supervised Direct (Training):** Training two networks independently using Kinetics-Sounds (34 classes).
3.2 Audio Visual Correspondence

**Supervised Direct (Testing):** Scalar product of two networks’ softmax output.
3.2 Audio Visual Correspondence

**Supervised Pretraining (Training):** Combine two trained networks, freeze the trained networks’ weights, then train the new fully connected layers.
## 3.2 Audio Visual Correspondence Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Flickr-SoundNet</th>
<th>Kinetics-Sounds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised direct</td>
<td>–</td>
<td>65%</td>
</tr>
<tr>
<td>Supervised pretraining</td>
<td>–</td>
<td>74%</td>
</tr>
<tr>
<td>$L^3$-Net</td>
<td>78%</td>
<td>74%</td>
</tr>
</tbody>
</table>
3.3 Audio Features

Environmental Sound Classification (ECS-50): 2000 audio clips, 5 seconds each, equally balanced between 50 classes. They extracted 10 Equally spaced 1 sec subclips.

Detection and Classification of Acoustic Scenes and Events (DCASE): 110 audio clips, 30 seconds each, with 10 classes. 60 subclips are extracted.
3.3 Audio Features

**L3-Net (Training):** Audio subnetwork of L3-Net which was pretrained on Flickr-SoundNet is used.

- Freeze weights
- Pretrained on Flickr
- Train SVM

Diagram showing the audio subnetwork with a max pool operation and dimensions 4*3*512 = 6144.
3.3 Audio Features

L3-Net (Testing): Audio subnetwork of L3-Net which was pretrained on Flickr-SoundNet is used.
3.3 Audio Features

L3-Net (random): Audio subnetwork of L3-Net which was initialized randomly is used.
### 3.3 Audio Features

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM-MFCC [26]</td>
<td>39.6%</td>
<td>RG [27]</td>
<td>69%</td>
</tr>
<tr>
<td>Random Forest [26]</td>
<td>44.3%</td>
<td>RNH [28]</td>
<td>77%</td>
</tr>
<tr>
<td>Piczak ConvNet [25]</td>
<td>64.5%</td>
<td>Ensemble [32]</td>
<td>78%</td>
</tr>
<tr>
<td>Ours random</td>
<td>62.5%</td>
<td>Ours random</td>
<td>85%</td>
</tr>
<tr>
<td>Ours</td>
<td>79.3%</td>
<td>Ours</td>
<td>93%</td>
</tr>
<tr>
<td>Human perf. [26]</td>
<td>81.3%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[Arandjelovic, Relja, and Andrew Zisserman. "Look, listen and learn." ICCV 2017]
3.4 Visual Features

**L3-Net (Training):** Visual subnetwork of L3-Net which was pretrained on Flickr-SoundNet is trained on ImageNet.
3.4 Visual Features

**L3-Net (random):** Visual subnetwork of L3-Net with random weights is trained on ImageNet.
3.4 Visual Features

<table>
<thead>
<tr>
<th>Method</th>
<th>Top 1 accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>18.3%</td>
</tr>
<tr>
<td>Pathak et al. [24]</td>
<td>22.3%</td>
</tr>
<tr>
<td>Krähenbühl et al. [16]</td>
<td>24.5%</td>
</tr>
<tr>
<td>Donahue et al. [7]</td>
<td>31.0%</td>
</tr>
<tr>
<td>Doersch et al. [6]</td>
<td>31.7%</td>
</tr>
<tr>
<td>Zhang et al. [36] (init: [16])</td>
<td>32.6%</td>
</tr>
<tr>
<td>Noroozi and Favaro [21]</td>
<td>34.7%</td>
</tr>
<tr>
<td>Ours random</td>
<td>12.9%</td>
</tr>
<tr>
<td>Ours</td>
<td>32.3%</td>
</tr>
</tbody>
</table>
3.5.1 Qualitative Analysis - Vision Features

For each particular unit in “Pool4”, test images with highest activation were chosen.
3.5.1 Vision Features on Kinetics-Sounds

[Arandjelovic, Relja, and Andrew Zisserman. "Look, listen and learn." ICCV 2017]
3.5.1 Vision Features on Kinetics-Sounds

[Arandjelovic, Relja, and Andrew Zisserman. "Look, listen and learn." ICCV 2017]
3.5.1 Vision Features on Flickr-SoundNet

[Arandjelovic, Relja, and Andrew Zisserman. "Look, listen and learn." ICCV 2017]
3.5.1 Audio Features on Kinetics-Sounds

[Arandjelovic, Relja, and Andrew Zisserman. "Look, listen and learn." ICCV 2017]
3.5.1 Audio Features on Flickr-SoundNet

[Arandjelovic, Relja, and Andrew Zisserman. "Look, listen and learn." ICCV 2017]
3.5.1 Audio Features on Flickr-SoundNet

[Arandjelovic, Relja, and Andrew Zisserman. "Look, listen and learn." ICCV 2017]
3.5.1 Audio Features on Flickr-SoundNet

[Arandjelovic, Relja, and Andrew Zisserman. "Look, listen and learn." ICCV 2017]
3.5.1 Audio Features on Kinetics-Sounds

[Arandjelovic, Relja, and Andrew Zisserman. "Look, listen and learn." ICCV 2017]
T-SNE on Audio

[Arandjelovic, Relja, and Andrew Zisserman. "Look, listen and learn." ICCV 2017]
T-SNE on Video

[Arandjelovic, Relja, and Andrew Zisserman. "Look, listen and learn." ICCV 2017]
Strength & Weakness

- **Strengths**
  - Introduces a new method to utilize audio to train video networks, and vice-versa
  - Lots of training data is available for self-supervised models

- **Weakness**
  - Same architecture for both subnetworks
  - Uses VGG architecture (why not at least ResNet? Maybe a little lazy because authors are in VGG group)
  - Not enough explanation on why their model works or fails on some tasks
  - Not exploiting temporal feature of video in visual subnetwork
  - Might not be able to learn finer details of an action (guitar picking v. power chords)
  - You can train anything quickly if you work for Google :(