The Sound of Pixels

H ZHAO, C GAN, A ROUDITCHENKO, C VONDRICK, J MCDERMOTT, A TORRALBA. EUROPEAN CONFERENCE ON COMPUTER VISION (ECCV) 2018
PRESENTED BY: JEFF HILDEBRANDT
Interactive Demo

Sound of Pixels Demo

http://sound-of-pixels.csail.mit.edu/
PixelPlayer - Overview

a) Input video

b) Audio-visual sounds source separation and localization

S(t)

l(x,y,t)

c) Estimated sound components for each location

y

x

d)...
e)
Technologies

- Sound source separation
- Learning visual-audio correspondence
- Self-supervised learning
Mix-and-Separate framework for Self-supervised Training

[Diagram showing the process of mixing and separating audio and video data.]
Mix-and-Separate framework for Self-Supervised Training

\[ S_{\text{mix}} = \sum_{n=1}^{N} S_n \]
\[ \hat{S}_n = f(S_{\text{mix}}, I_n) \]

Binary Masks - Per-pixel sigmoid cross entropy loss
\[ M_n(u, v) = [S_n(u, v) \geq S_m(u, v)], \quad \forall m = (1, ..., N) \]

Ratio Masks - Per-pixel L1 loss
\[ M_n(u, v) = \frac{S_n(u, v)}{S_{\text{mix}}(u, v)} \]
MUSIC Dataset

685 untrimmed videos of musical solos and duets 11 instrument categories
MUSIC Dataset
Audio Data Processing

- 11kHz sub-sampled audio signals
  - Highest signal frequency is 5.5kHz
- 6 second segments
- STFT with window size 1022 and hop length of 256
- 512x256 Time-Frequency

Output from the model:
- Use an inverse sampling to convert the mask back to linear frequency
- Perform inverse STFT to get the recovered signal
Model – Video Analysis Network

- Modified ResNet-18
- Removed last average pooling layer
- Removed last fc layer
- Removed stride for last residual block
- Last residual block has a dilation of 2
- Added a 3x3 convolutional layer with K output channels
- T frames with size 224x224x3 as input
- Outputs a feature of size K after spatiotemporal max pooling
Model – Audio Analysis Network

• Modified U-Net
• 7 convolutions (down-convolutions)
• 7 de-convolutions (up-convolutions)
• Skip connections between the convolutions and de-convolutions
• Input audio spectrogram with size 256x256x1
• Outputs K feature maps of size 256x256xk
U-Net

Model – Audio Synthesizer

• Accepts outputs from video and audio analysis networks
• Fuses outputs
• Outputs a mask
• Linear layer
  • $K$ weights + 1 bias
• Best model takes 3 frames as visual input and uses $K=16$ feature channels

\[
\sum_{k=1}^{K} \alpha_k i_k(x, y) s_k + \beta_0
\]
Implementation Details

- 500 videos for training
  - Solos and duets
- 130 videos for validation
  - Only solos
- 84 videos for testing
  - Only duets
- Silent videos
  - Background images from ADE dataset

- Use N=2 videos
- SGD optimizer
  - Momentum 0.9
  - Learning rate 0.0001
Implementation Details
Results Format

\[ S_{\text{target}} \quad \text{Source to Distortion} \quad SDR := 10 \log_{10} \frac{\| S_{\text{target}} \|^2}{\| e_{\text{interf}} + e_{\text{noise}} + e_{\text{artif}} \|^2} \]

\[ e_{\text{interf}} \]

\[ e_{\text{noise}} \quad \text{Source to Interferences} \quad \text{SIR} := 10 \log_{10} \frac{\| S_{\text{target}} \|^2}{\| e_{\text{interf}} \|^2} \]

\[ e_{\text{artif}} \quad \text{Source to Artifacts} \quad \text{SAR} := 10 \log_{10} \frac{\| S_{\text{target}} + e_{\text{interf}} + e_{\text{noise}} \|^2}{\| e_{\text{artif}} \|^2} \]
Results

<table>
<thead>
<tr>
<th></th>
<th>NMF</th>
<th>DeepConvSep</th>
<th>Spectral Regression</th>
<th>Ratio Mask Linear scale</th>
<th>Ratio Mask Log scale</th>
<th>Binary Mask Linear scale</th>
<th>Binary Mask Log scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSDR</td>
<td>3.14</td>
<td>6.12</td>
<td>5.12</td>
<td>6.67</td>
<td>8.56</td>
<td>6.94</td>
<td>8.87</td>
</tr>
<tr>
<td>SIR</td>
<td>6.70</td>
<td>8.38</td>
<td>7.72</td>
<td>12.85</td>
<td>13.75</td>
<td>12.87</td>
<td>15.02</td>
</tr>
<tr>
<td>SAR</td>
<td>10.10</td>
<td>11.02</td>
<td>10.43</td>
<td>13.87</td>
<td>14.19</td>
<td>11.12</td>
<td>12.28</td>
</tr>
</tbody>
</table>

- NSDR - Normalized Signal-to-Distortion Ratio
- SIR - Signal-to-Interference Ratio
- SAR - Signal-to-Artifact Ratio
Sound Localization and Clustering of Sounds

Which pixels are making sounds?

What sounds do these pixels make?
Visual-audio corresponding activations
Discriminative channel activations

<table>
<thead>
<tr>
<th>IoU Threshold</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy(%)</td>
<td>66.10</td>
<td>47.92</td>
<td>32.43</td>
</tr>
</tbody>
</table>
Subjective Evaluations - Sound

Which sound do you hear?

<table>
<thead>
<tr>
<th>Model</th>
<th>Correct (%)</th>
<th>Wrong (%)</th>
<th>Both (%)</th>
<th>None (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMF</td>
<td>45.70</td>
<td>15.23</td>
<td>21.35</td>
<td>17.71</td>
</tr>
<tr>
<td>Spectral Regression</td>
<td>18.23</td>
<td>15.36</td>
<td>64.45</td>
<td>1.95</td>
</tr>
<tr>
<td>Ratio Mask</td>
<td>39.19</td>
<td>19.53</td>
<td>27.73</td>
<td>13.54</td>
</tr>
<tr>
<td>Binary Mask</td>
<td><strong>59.11</strong></td>
<td>11.59</td>
<td>18.10</td>
<td>11.20</td>
</tr>
<tr>
<td>Ground Truth Solo</td>
<td>70.31</td>
<td>16.02</td>
<td>7.68</td>
<td>5.99</td>
</tr>
</tbody>
</table>
Subjective Evaluations – Visual-Sound Correspondence

Is the sound coming from this pixel?

<table>
<thead>
<tr>
<th>Model</th>
<th>Yes(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spectral Regression</td>
<td>39.06</td>
</tr>
<tr>
<td>Ratio Mask</td>
<td>54.68</td>
</tr>
<tr>
<td>Binary Mask</td>
<td>67.58</td>
</tr>
</tbody>
</table>
Thank you!

Questions?