WHAT MAKES THE SOUND?: A DUAL-MODALITY INTERACTING NETWORK FOR AUDIO-VISUAL EVENT LOCALIZATION

Janani Ramaswamy
Dept. of CS & E, IIT Madras, Chennai-600036, India
Can’t machines mimic humans in using both audio and video for decision making?
CHALLENGES
➢ Audio may not always be in perfect sync with the video

➢ Presence of ambient sound like breeze

➢ Object making the sound being momentarily occluded in the video

➢ Obtaining the semantics is less direct in case of audio\(^1\).

Supervised event localization:
Training: event label given for every segment
Testing: predict event category for every segment

Weakly-Supervised event localization:
Training: event label given for whole video
Testing: predict event category for every segment
APPLICATIONS
➢ Audio-based video captioning
➢ Audio-based video segmentation
➢ Surveillance
➢ Enhanced scene understanding
## EXISTING WORKS

<table>
<thead>
<tr>
<th>Tian et al. ECCV 2018 ¹</th>
<th>Wu et al. ICCV 2019 ³</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Audio-visual event localization in unconstrained videos</td>
<td>- Dual Attention Matching (DAM)</td>
</tr>
<tr>
<td>- Audio-Visual Event (AVE) dataset</td>
<td>- Encodes temporal co-occurrence between auditory and visual signals</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lin et al. ICASSP 2019 ²</th>
<th>Ramaswamy &amp; Das WACV 2020 ⁴</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Audio-Visual seq2seq dual n/w (AVSDN)</td>
<td>- Spatial &amp; Segment-wise attention using two novel blocks</td>
</tr>
<tr>
<td>- learns global and local event info in seq2seq manner</td>
<td>- A novel loss function for unsupervised sound localization</td>
</tr>
</tbody>
</table>

MAJOR CONTRIBUTIONS

Audio-Visual Interacting Network (AVIN) for fully & weakly supervised audio-visual event localization

A novel audio-visual fusion that captures the inter and intra modality interactions using local and global information from the two modalities

Our method significantly outperforms the existing state-of-the-art methods
PROPOSED ARCHITECTURE
Audio-Visual Interacting Network (AVIN)

• Feature Extraction: Let $F_t^a \in \mathbb{R}^{d_a}$ and $F_t^v \in \mathbb{R}^{d_v}$ denote the audio and visual features extracted using CNNs. Here, $d_a$ and $d_v$ refer to the dimension of audio and visual features respectively.

• Modeling temporal dependency: The features $\{F_t^a, F_t^v\}_{t=1}^{\mathcal{T}}$ extracted from the CNNs are then fed to two different LSTMs, the result of which is denoted as $\{h_t^a, h_t^v\}_{t=1}^{\mathcal{T}}$.

• $\mathcal{T}$ here denotes the number of non-overlapping segments ($= 10$ in our case) that each video is split into.
CAPTURING BILINEAR INTERACTIONS

[Diagram showing a network of neural network components including CNN, LSTM, dropout, sum pooling, normalization, and event category inputs and outputs.]
Bilinear Pooling for audio-visual fusion

• Consider a multi-modal bilinear model:
  \[ \tilde{z}_t = F_t^v W_i F_t^a \]
  \[ \text{----------- (1)} \]
  where, \( W_i \in \mathbb{R}^{d_v \times d_a} \) is the projection matrix and \( \tilde{z}_t \) is a scalar.

• To get a \( p \)-dimensional output, we use \( W = [W_1, ..., W_p] \in \mathbb{R}^{d_v \times d_a \times p} \)

• But this leads to a large number of parameters and high computational cost.

• Multi-modal Factorized Bilinear (MFB) Pooling\(^1\) factorizes \( W \) into two low-rank matrices:
  \[ \tilde{z}_t = \text{Sumpooling}(U^T F_t^v \circ V^T F_t^a, q) \]
  \[ \text{----------- (2)} \]

• Applying power and L2 normalization:
  \[ z'_t = \text{sign}(\tilde{z}_t)|\tilde{z}_t|^{0.5}; z_t = z'_t/||z'_t|| \]
  \[ \text{----------- (3)} \]

Where, \( U \in \mathbb{R}^{d_v \times (qp)} \) and \( V \in \mathbb{R}^{d_a \times (qp)} \) are the two low rank matrices.
\( \circ \) refers to the Hadamard product and \( q \) represents the latent dimensionality.

Capturing inter and intra modality interactions
Capturing inter and intra modality interactions

• To get a better idea about the amount of synchronization present between the two modalities, the global information also needs to be considered.
• We use self and collaborative attention\(^1\) to capture intra and inter modality interactions.
• **Intra-modality interactions:**

\[
s_t^a = \text{Softmax} (h_t^a \odot \overline{h}_{ave}^a) \otimes h_t^a \tag{4}
\]
\[
s_t^v = \text{Softmax} (h_t^v \odot \overline{h}_{ave}^v) \otimes h_t^v \tag{5}
\]

• **Inter-modality interactions:**

\[
c_t^a = \text{Softmax} (h_t^a \odot \overline{s}_{ave}^v) \otimes h_t^a \tag{6}
\]
\[
c_t^v = \text{Softmax} (h_t^v \odot \overline{s}_{ave}^a) \otimes h_t^v \tag{7}
\]

where,

- \(h_t^a, h_t^v\) - temporally encoded features from LSTMs
- \(\overline{h}_{ave}^a, \overline{h}_{ave}^v\) - outputs of mean pooling applied on \(h_t^a, h_t^v\)
- \(s_t^a, s_t^v\) - features encoded with **intra**-modality interactions
- \(c_t^a, c_t^v\) - features encoded with **inter**-modality interactions
- \(\overline{s}_{ave}^a, \overline{s}_{ave}^v\) - outputs of mean pooling applied on \(s_t^a, s_t^v\)

\(\odot\) - dot product
\(\otimes\) - element-wise multiplication

\(^1\) Zhang et al., Scan: Self-and-collaborative attention network for video person re-identification, TIP 2019.
DATASET USED

Audio-Visual Event (AVE) Dataset\(^1\)

- 4143 videos (min 2s long event; max 10s long event)
- 28 event categories
- Minimum of 60 and maximum of 188 videos in each category
- Labels available video-wise as well as segment-wise (i.e., temporally labeled) with audio-visual event boundaries.

### RESULTS (PERFORMANCE COMPARISON IN %)

<table>
<thead>
<tr>
<th>Method</th>
<th>Sup. Acc.</th>
<th>W-Sup. Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio</td>
<td>62.3</td>
<td>57.0</td>
</tr>
<tr>
<td>Visual</td>
<td>57.4</td>
<td>53.8</td>
</tr>
<tr>
<td>AVE(^1)</td>
<td>72.7</td>
<td>66.7</td>
</tr>
<tr>
<td>AVSDN(^2)</td>
<td>72.8</td>
<td>66.5</td>
</tr>
<tr>
<td>DAM(^3)</td>
<td>74.5</td>
<td>-</td>
</tr>
<tr>
<td>Ramaswamy &amp; Das(^4)</td>
<td>74.8</td>
<td>68.9</td>
</tr>
<tr>
<td>AVIN (Ours: Aud + Vis)</td>
<td>75.2</td>
<td>69.4</td>
</tr>
</tbody>
</table>

# RESULTS (DIFFERENT FUSION STRATEGIES)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Element-wise multiplication</td>
<td>60.3</td>
<td>55.1</td>
</tr>
<tr>
<td>Element-wise addition</td>
<td>63.4</td>
<td>58.2</td>
</tr>
<tr>
<td>Concatenation + FC</td>
<td>65.7</td>
<td>60.3</td>
</tr>
<tr>
<td>AVIN (Ours)</td>
<td>75.2</td>
<td>69.4</td>
</tr>
</tbody>
</table>
### ABLATION STUDY

<table>
<thead>
<tr>
<th>Model</th>
<th>Sup. Acc.</th>
<th>W-Sup. Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Only LSTM</td>
<td>70.1</td>
<td>63.8</td>
</tr>
<tr>
<td>Only MFB(^1)</td>
<td>71.4</td>
<td>66.7</td>
</tr>
<tr>
<td>LSTM + intra-mod</td>
<td>71.2</td>
<td>65.4</td>
</tr>
<tr>
<td>LSTM + intra + inter-mod</td>
<td>73.5</td>
<td>67.9</td>
</tr>
<tr>
<td>LSTM + MFB + intra+ inter-mod</td>
<td>75.2</td>
<td>69.4</td>
</tr>
</tbody>
</table>

Bar chart depicting accuracies of a few selected event categories for supervised event localization task.
Output of a few segments shown for our proposed method of supervised event localization, given an input video.
THANK YOU!

JANANI RAMASWAMY
(Research Scholar, IIT Madras)

Visualization and Perception Lab – www.cse.iitm.ac.in/~vplab