AUDIO-VISUAL FUSION AND CONDITIONING WITH NEURAL NETWORKS FOR EVENT RECOGNITION

Mathilde Brousmiche $^{1,2}$  Jean Rouat $^2$  Stéphane Dupont $^1$

$^1$Numediart Institute
University of Mons
Belgium

$^2$Necotis Lab
University of Sherbrooke
Canada

MLSP, October 2019
Multimodality
Problem setting: Audio-visual Event Classification

Subset of kinetics$^1$:

- blowing_nose
- clapping
- crying
- finger_snapping
- playing_drums
- playing_guitar
- sneezing
- using_computer
- whistling
- yawning

Problem setting: Audio-visual Event Classification

Visual

Playing guitar

M. Brousmiche
Problem setting: Audio-visual Event Classification

Visual

Audio

Playing guitar

Playing guitar
Problem setting: Audio-visual Event Classification
Problem setting: Audio-visual Event Classification

- **Fusion?**
- **Conditioning?**

Visual

Audio

Playing guitar
Techniques of fusion

- **Concatenation**
- **Element-wise addition**
- **Multimodal Compact Bilinear pooling (MCB)**

---

Fusion levels: Early fusion

Visual feature extractor (DenseNet)

Audio feature extractor (CNN)

Fusion Block

Playing guitar
Fusion levels: Middle fusion

Audio-visual Fusion

Visual feature extractor (DenseNet)

Audio feature extractor (CNN)

Fusion Block

Playing guitar

M. Brousmiche

AUDIO-VISUAL FUSION AND CONDITIONING WITH NEURAL NETWORKS FOR EVENT RECOGNITION
Fusion levels: Late fusion

- Visual feature extractor (DenseNet)
- Audio feature extractor (CNN)
- Fusion Block

Playing guitar
Audio-visual Fusion efficiency

![Bar chart showing accuracy in % for different conditions: Image, Sound, Concat - Early, Concat - Middle, Add - Middle, Add - Late, MCB - Middle. The chart compares unimodal classification (blue) and multimodal classification (orange).]
What is conditioning?
What is conditioning?
Modalities conditioning with a attention model\(^2\)

Proposal: Visual feature map modulation with audio information
Proposal: Visual feature map modulation with audio information
Feature-wise Linear Modulation (FiLM)

\( \gamma_{i,c} \) and \( \beta_{i,c} \) modulate the activations \( F_{i,c} \):

\[
FiLM(F_{i,c} | \gamma_{i,c}, \beta_{i,c}) = \gamma_{i,c} F_{i,c} + \beta_{i,c}
\]

where

\[
\gamma_{i,c} = f_c(x_i) \quad \beta_{i,c} = h_c(x_i)
\]

\( f \) and \( h \) can be arbitrary functions

---

## Contribution of FiLM in audio-visual event classification

<table>
<thead>
<tr>
<th></th>
<th>Image</th>
<th>Sound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without FiLM modula</td>
<td>$61.00 \pm 5.11$</td>
<td>$66.67 \pm 4.60$</td>
</tr>
<tr>
<td>With FiLM modula</td>
<td>$75.75 \pm 5.35$</td>
<td>$75.75 \pm 3.14$</td>
</tr>
</tbody>
</table>
Better embedding clustering with FiLM

(a) Without FiLM  (b) With FiLM
Better embedding clustering with FiLM
Better embedding clustering with FiLM

(a) Without FiLM

(b) With FiLM

- blowing_nose
- clapping
- crying
- finger_snapping
- playing_drums
- playing_guitar
- sneezing
- using_computer
- whistling
- yawning
Better embedding clustering with FiLM

(a) Without FiLM  (b) With FiLM

• blowing_nose  • playing_drums  • using_computer
• clapping  • playing_guitar  • whistling
• crying  • sneezing  • yawning

M. Brousmiche

UMons, UdeS

AUDIO-VISUAL FUSION AND CONDITIONING WITH NEURAL NETWORKS FOR EVENT RECOGNITION
Conclusion

- Relevant information for event recognition exists both in visual and audio modalities.
- Exploiting both audio and visual modalities through fusion or conditioning improves event recognition performance.
- The use of FiLM layers allows exploiting both audio and visual modalities without an explicit implementation of the fusion.
Future Work

- Test another conditioning method based on multimodal Long Short-Term Memory (LSTM) neural networks
- Analyze the robustness of all methods in the presence of noise as well as in the absence of one modality.
<table>
<thead>
<tr>
<th>Introduction</th>
<th>Audio-visual Fusion</th>
<th>Audio-visual Conditioning</th>
<th>Conclusion</th>
</tr>
</thead>
</table>

Thank you!