Implicit Fusion by Joint Audiovisual Training for Emotion Recognition in Mono Modality

Jing Han¹, Zixing Zhang², Zhao Ren¹, Björn Schuller¹,²

1. ELHW, University of Augsburg, Germany
2. GLAM, Imperial College London, UK
Motivation:
- mono or multi ?
- multi > mono, true always?
- MTL with auxiliary task(s)

*Issue:* how to exploit info. across modalities

*we propose implicit fusion*
- auxiliary modality for training
- mono modality for inference

Jing Han
Implicit Fusion by Joint Audiovisual Training for Emotion Recognition in Mono Modality

(a) early fusion  (b) late fusion  (c) model-level fusion  (d) implicit fusion

© concatenation  ✔ voting  □ shared layers  ℹ specific layers

Jing Han
Implicit Fusion by Joint Audiovisual Training for Emotion Recognition in Mono Modality

multi-task learning

implicit fusion

auxiliary modalities help train the shared networks
During training:

**Joint training:**

*one batch per modality*

*(paired/unpaired)*

**Loss to optimise:**

**Audio:** \[ \mathcal{L} = \mathcal{L}_a + \alpha \mathcal{L}_v \]

**Video:** \[ \mathcal{L} = \mathcal{L}_v + \alpha \mathcal{L}_a \]
During inference:

- simply remove auxiliary modality
- and evaluate in mono modality
Datasets and features:

- **One-Minute Gradual-Emotional (OMG-Emotion) for 7-class classification**
  
  - train: 2440, dev: 617, test: 2229
  - audio: 88 eGeMAPS
  - video: VGGFace, fc7-4096

- **RECOLA with 27 audiovisual recordings for arousal and valence regression**
  
  - train/dev/test: 67.5K
  - audio: 88 eGeMAPS
  - video: 632 geometric features
Implicit Fusion by Joint Audiovisual Training for Emotion Recognition in Mono Modality

Experimental settings:

- 2/4 GRU layers and 100 cells per layer for OMG/RECOLA
- Adam with initial learning rate .001
- Minibatch size 128
- Grid search [0,1] with a step of .1 to optimise $\alpha$
- Results in terms of F1-score/CCC
Results on OMG-emotion

<table>
<thead>
<tr>
<th>Method</th>
<th>F1 [%]</th>
<th>audio</th>
<th>video</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM [23]</td>
<td>33.0</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>CNN [23]</td>
<td>—</td>
<td>37.0</td>
<td>—</td>
</tr>
<tr>
<td>RNN (baseline)</td>
<td>36.5</td>
<td>37.9</td>
<td>—</td>
</tr>
<tr>
<td>RNN (implicit fusion)</td>
<td>40.2</td>
<td>42.1</td>
<td>—</td>
</tr>
</tbody>
</table>
## Results on RECOLA

<table>
<thead>
<tr>
<th>CCC</th>
<th>arousal</th>
<th></th>
<th></th>
<th>valence</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>audio</td>
<td>video</td>
<td>audio</td>
<td>video</td>
<td>audio</td>
<td>video</td>
</tr>
<tr>
<td></td>
<td>dev.</td>
<td>test</td>
<td>dev.</td>
<td>test</td>
<td>dev.</td>
<td>test</td>
</tr>
<tr>
<td>DNN+Curriculum learning [29]</td>
<td>.687</td>
<td>.591</td>
<td>.394</td>
<td>.267</td>
<td>.159</td>
<td>.174</td>
</tr>
<tr>
<td>MTL (RE based) [30]</td>
<td>.788</td>
<td>.629</td>
<td>.502</td>
<td>.324</td>
<td>.519</td>
<td>.331</td>
</tr>
<tr>
<td>MTL (PU based) [12]</td>
<td>.803</td>
<td>.654</td>
<td>.508</td>
<td>.327</td>
<td>.506</td>
<td>.416</td>
</tr>
<tr>
<td>DDAT (RE based) [22]</td>
<td>.807</td>
<td>.694</td>
<td>.544</td>
<td>.400</td>
<td>.508</td>
<td>.422</td>
</tr>
<tr>
<td>DDAT (PU based) [22]</td>
<td>.811</td>
<td>.664</td>
<td>.513</td>
<td>.397</td>
<td>.498</td>
<td>.407</td>
</tr>
<tr>
<td>RNN (baseline)</td>
<td>.766</td>
<td>.605</td>
<td>.499</td>
<td>.399</td>
<td>.504</td>
<td>.381</td>
</tr>
<tr>
<td>RNN (implicit fusion)</td>
<td>.769</td>
<td>.611</td>
<td>.515</td>
<td>.413</td>
<td>.513</td>
<td>.395</td>
</tr>
</tbody>
</table>

Jing Han
## Results on RECOLA

<table>
<thead>
<tr>
<th>CCC</th>
<th>arousal audio test</th>
<th>video dev. test</th>
<th>valence audio test</th>
<th>video dev. test</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNN+Curriculum learning [29]</td>
<td>.687 .591</td>
<td>.394 .267</td>
<td>.159 .174</td>
<td>.300 .269</td>
</tr>
<tr>
<td>MTL (RE based) [30]</td>
<td>.788 .629</td>
<td>.502 .324</td>
<td>.519 .331</td>
<td>.632 .488</td>
</tr>
<tr>
<td>MTL (PU based) [12]</td>
<td>.803 .654</td>
<td>.508 .327</td>
<td>.506 .416</td>
<td>.643 .452</td>
</tr>
<tr>
<td>DDAT (RE based) [22]</td>
<td>.807 <strong>.694</strong></td>
<td>.544 .400</td>
<td>.508 <strong>.422</strong></td>
<td>.639 .471</td>
</tr>
<tr>
<td>DDAT (PU based) [22]</td>
<td>.811 .664</td>
<td>.513 .397</td>
<td>.498 .407</td>
<td>.632 .501</td>
</tr>
<tr>
<td>RNN (baseline)</td>
<td><strong>.766</strong> .605</td>
<td><strong>.499</strong> .399</td>
<td><strong>.504</strong> .381</td>
<td><strong>.619</strong> .529</td>
</tr>
<tr>
<td>RNN (implicit fusion)</td>
<td><strong>.769</strong> .611</td>
<td><strong>.515</strong> <strong>.413</strong></td>
<td>.513 .395</td>
<td>.622 .527</td>
</tr>
</tbody>
</table>

- for **audio**, improvement for valence prediction is more
- for **video**, improvement for arousal prediction is more

Jing Han
Implicit Fusion by Joint Audiovisual Training for Emotion Recognition in Mono Modality

Conclusion:
- model targeted at mono-modality scenario
- audio and video knowledge explored during training implicitly
- loss from auxiliary modality to advance the learning
- strengths: free from alignment, unpaired data

Future work:
- across other modalities beyond audio and visual
- implicit fusion + DDAT
- maybe DANN

Take home message:
- lack of data? auxiliary modality might help!