Speaker and Emotion Recognition of TV-Series Data Using Multimodal and Multitask Deep Learning

Sashi Novitasari¹, Quoc Truong Do¹, Sakriani Sakti¹,³, Dessi Lestari², Satoshi Nakamura¹,³
¹ Graduate School of Information Science, Nara Institute of Science and Technology
² Department of Informatics, Bandung Institute of Technology
³ RIKEN AIP
¹{sashi.novitasari.si3, do.truong.dj3, ssakti, s-nakamura}@is.naist.jp
²{dessipuji}@informatika.org
I. Introduction

- Real-life communication involves linguistic and paralinguistic aspects
- Multimodal and multitask recognition of non-verbal aspects of speech
- Recognition of speech’s speaker and emotion from emotion-rich data
- Previous works:
  - Multimodal or multitask emotion-speaker recognition (not integrated)
    (Tang et al., 2016; Tian et al., 2016; Vallet et al., 2013)
II. Data

- **TV-series data** → expressive conversation
  - Video graphic: Facial features
  - Audio: Acoustic features
  - Subtitle: Lexical features

- English

- Utterance-level annotation
  - Speaker: 57 names
  - Emotion - valence: 3 classes (negative - neutral - positive)
  - Emotion - arousal: 3 classes (negative - neutral - positive)
III. Model Architectures

- Multilayer perceptron models (5 layers)
- Multimodal classification
- Multitask classification
III. Model Architectures

Multimodal Classification

2 evaluated approaches:

a. Features concatenation

b. Features hierarchical fusion
III. Model Architectures

Multitask Classification

Perform classification on several tasks at once.
IV. Features

1. Acoustic (main)
   ○ INTERSPEECH 2010 feature conf.
   ○ openSMILE toolkit (Eyben et al., 2010)

2. Lexical
   ○ Word-vectors average
   ○ Pre-trained Google Word2Vec (Mikolov et al., 2013)

3. Facial
   ○ Facial contours and angles
   ○ openFace toolkit (Baltrusaitis et al., 2016)
V. Experiment
V. Experiment

- **Train set:** 2460 utterances
  - Speaker: 57 speaker (imbalanced)
  - Valence: Negative 31%, Neutral 60%, Positive 9%
  - Arousal: Negative 4%, Neutral 75%, Positive 21%

- **Evaluated** on 300 utterances
  - Speaker: 10 speaker, 30 samples each
  - Valence: Negative 32%, Neutral 57%, Positive 11%
  - Arousal: Negative 1%, Neutral 78%, Positive 21%

- Compared performance of unimodal, multimodal, single-task, and multitask models

- Evaluated based on F1-score(%) on evaluation set
V. Experiment
Result
V. Experiment

Result: Speaker

F1-scores (%) on evaluation set

*Multimodal approaches
U - Unimodal
C - Features concatenation
H - Features hierarchical fusion

Feature types
A - Acoustic
F - Facial
L - Lexical
V. Experiment

**Result: Emotion**

F1-score (%) on evaluation set

*Multimodal approaches*
- U - Unimodal
- C - Features concatenation
- H - Features hierarchical fusion

*Feature types*
- A - Acoustic
- F - Facial
- L - Lexical
V. Experiment

Result Summary

*Multimodal approaches
U - Unimodal
C - Features concatenation
H - Features hierarchical fusion

Feature types
A - Acoustic
F - Facial
L - Lexical

<table>
<thead>
<tr>
<th></th>
<th>Speaker</th>
<th>Emotion (average)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F1-Score %</td>
<td>F1-Score %</td>
</tr>
<tr>
<td>U: A</td>
<td>Single-task: 62.77</td>
<td>Single-task: 49.45</td>
</tr>
<tr>
<td></td>
<td>Multitask: 60.13</td>
<td>Multitask: 52.96</td>
</tr>
<tr>
<td>C: A+F</td>
<td>Single-task: 56.73</td>
<td>Single-task: 50.62</td>
</tr>
<tr>
<td></td>
<td>Multitask: 58.51</td>
<td>Multitask: 51.25</td>
</tr>
<tr>
<td>C: A+F+L</td>
<td>Single-task: 56.23</td>
<td>Single-task: 50.88</td>
</tr>
<tr>
<td></td>
<td>Multitask: 61.01</td>
<td>Multitask: 50.94</td>
</tr>
<tr>
<td></td>
<td>Multitask: 52.71</td>
<td>Multitask: 49.98</td>
</tr>
</tbody>
</table>
VI. Conclusion

- We constructed the multimodal and multitask speaker-emotion recognition model by using deep learning and TV-series data.

- Multitask model able to outperform single-task model, especially when recognizing emotion by using acoustic features only.

- Multimodal-multitask model did not result in a significant improvement (larger data might be needed).
Thank You