Unsupervised domain adaptation for preference learning based speech emotion recognition

- Abinay Reddy Naini, Mary A. Kohler, Carlos Busso

Presentation by Abinay Reddy Naini
Emotion recognition is critical for the Intelligence Community (IC)

- Analyze massive amount of information available through media domains
- Identify and preselect segments with potentially threatening behaviors
Expression of Emotion

- Categorical labels
  - Anger, happiness, sadness, neutral

- Dimensional or attribute-based labels
  - Valence (negative versus positive)
  - Arousal (calm versus active)
  - Dominance (weak versus strong)
  - More accurate emotion descriptors (intensity)
Ordinal Representation of Emotion

- **Thesis: emotions are intrinsically ordinal (relative)**
  - The benefits of representing them that way are many!
  - This thesis is supported by theoretical arguments across disciplines and empirical evidence in Affective Computing.


Preference learning formulation

- Preference learning

<table>
<thead>
<tr>
<th>Sentence #1</th>
<th>Preference learning Based SER</th>
<th>Sentence #2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Arousal (Sentence #1 &gt;&gt; Sentence #2)</td>
<td>Valence (Sentence #2 &gt;&gt; Sentence #1)</td>
</tr>
<tr>
<td></td>
<td>Dominance (Sentence #2 &gt;&gt; Sentence #1)</td>
<td></td>
</tr>
</tbody>
</table>

- Why preference learning?
  - Humans are better at relative comparisons than absolute values
  - Appealing to Emotional Retrieval tasks
  - Better use of training data
    - $N(N-1)/2$ potential pairs

Getting preference labels?
The goal is to define trends in the evaluations.

In the above example, there are 15 preferences for sentence one, 2 preferences for sentence two, and 7 draws.
RankNet (Previous works)

Ideal probabilities $\hat{P}_{ij}$ is set according to the preference in pairs of samples.

- $\hat{P}_{ij} = 0$ if $j \gg i$
- $\hat{P}_{ij} = 1$ if $i \gg j$

$$C = -\hat{P}_{ij}\log P_{ij} - (1 - \hat{P}_{ij})\log(1 - P_{ij})$$

Sample sentences $i, j$, with features $\Phi_i, \Phi_j$

- We can implement function $f()$ with arbitrary architectures
- Focus on generalization
  - Train on one domain and test on another
- We consider two alternative and complementary domain adaptation schemes
  - Ladder networks (feature reconstruction)
  - Adversarial domain adaptation (feature representation)
Chunk based segmentation

Feature Extraction
- e.g., spectrogram, LLDS, waveform

Dynamic Chunk Segmentation

Chunk-level Feature Representation
- e.g., LSTM, CNN, Functional

Sentence-level Temporal Aggregation
- e.g., NonAtten, Gate/Vec, RNN-AttenVec, Self-AttenVec

Raw Feature Map (X)

Wav2Vec 2.0

2-LSTM layers

RNN-AttenVec

\[ v = \sum_{t=1}^{C} \alpha_t \overline{h_t} \]

\[ z = \tanh(W[v; \overline{h_C}]) \]
Adversarial Domain Adaptation Model

Task classifier loss = $L_t$

Domain classifier loss = $L_d$

The goal is to recover a clean version of the encoder while obtaining task-specific encoded features.

- Source domain: Both task classifier loss along with reconstruction loss.
- Target domain: Only reconstruction loss.

Total cost:

\[ C = C_c + \lambda t \sum l C_d^{(l)} \]
Proposed Architectures

Proposed Architectures

Proposed Architectures

Proposed Architectures

Data preparation

- **MSP-Podcast v1.10 (Source domain)**
  - Recordings are annotated for emotional attribute labels (arousal, valence, dominance)
  - We have used ~420k pairs of samples from MSP-Podcast training set.

- **MSP-IMPROV (Target domain)**
  - Recordings are annotated for emotional attribute labels (arousal, valence, dominance) like MSP-Podcast.
  - Data from the first three sessions are used as the test set, remaining sessions are reserved for adaptation.
  - We sample an equal number of pairs (~420k) for adaptation.
Feature extraction

- Wav2vec2-large-robust\(^1\)
  - wav2vec2-large feature representation (1024) using pre-trained Wav2vec2.0 large model from the HuggingFace library.
  - Then, we prune the top 12 transformer blocks, and fine-tuned the model using the MSP-Podcast corpus.
- extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS)\(^2\)
  - We also present a baseline using the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS), which includes 88 acoustic features.

---

Kendall’s Tau correlation coefficient (KT)

- If \((x_1, y_1) \ldots (x_n, y_n)\) be a set of observations
- \((x_i, x_j), (y_i, y_j)\) are said to be concordant if the sort order agrees.

\[
KT = \frac{(\text{Number of concordant pairs}) - (\text{Number of discordant pairs})}{\binom{n}{2}}
\]

- For testing: 200 utterance are sampled randomly. This process is repeated 20 times, then mean and SD of the result are reported.
## Results

**KT: Kendall’s Tau correlation coefficient**

<table>
<thead>
<tr>
<th>Method</th>
<th>Labeled</th>
<th>Un-labeled</th>
<th>Arousal</th>
<th>Valence</th>
<th>Dominance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>KT</td>
<td>Accuracy</td>
<td>KT</td>
</tr>
<tr>
<td>eGeMAPS+2FC</td>
<td>MSP-PODCAST</td>
<td>-</td>
<td>0.372</td>
<td>64.3</td>
<td>0.208</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.316</td>
</tr>
<tr>
<td>CBS+2FC</td>
<td>MSP-PODCAST</td>
<td>-</td>
<td>0.417</td>
<td>71.9</td>
<td>0.258</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.392</td>
</tr>
<tr>
<td>CBS+LN</td>
<td>MSP-PODCAST</td>
<td>MSP-IMPROV</td>
<td>0.484</td>
<td>79.2</td>
<td>0.313</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.447</td>
</tr>
<tr>
<td>CBS+DA</td>
<td>MSP-PODCAST</td>
<td>MSP-IMPROV</td>
<td>0.462</td>
<td>78.5</td>
<td>0.301</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.442</td>
</tr>
<tr>
<td>CBS+LN+DA</td>
<td>MSP-PODCAST</td>
<td>MSP-IMPROV</td>
<td><strong>0.506</strong></td>
<td><strong>80.6</strong></td>
<td><strong>0.312</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>0.461</strong></td>
</tr>
</tbody>
</table>

**Observations:**
- Lower performance without domain adaptation
## Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Labeled</th>
<th>Un-labeled</th>
<th>Arousal</th>
<th>Valence</th>
<th>Dominance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>KT</td>
<td>Accuracy</td>
<td>KT</td>
</tr>
<tr>
<td>eGeMAPS+2FC</td>
<td>MSP-PODCAST</td>
<td>-</td>
<td>0.372</td>
<td>64.3</td>
<td>0.208</td>
</tr>
<tr>
<td>CBS+2FC</td>
<td>MSP-PODCAST</td>
<td>-</td>
<td>0.417</td>
<td>71.9</td>
<td>0.258</td>
</tr>
<tr>
<td>CBS+LN</td>
<td>MSP-PODCAST</td>
<td>MSP-IMPROV</td>
<td>0.484</td>
<td>79.2</td>
<td>0.313</td>
</tr>
<tr>
<td>CBS+DA</td>
<td>MSP-PODCAST</td>
<td>MSP-IMPROV</td>
<td>0.462</td>
<td>78.5</td>
<td>0.301</td>
</tr>
<tr>
<td>CBS+LN+DA</td>
<td>MSP-PODCAST</td>
<td>MSP-IMPROV</td>
<td>0.506</td>
<td>80.6</td>
<td>0.312</td>
</tr>
</tbody>
</table>

### Observations:
- Both adaptation methods lead to improvements
- Best performance achieved by combining ladder network and domain adaptation
## Results

**KT: Kendall’s Tau correlation coefficient**

<table>
<thead>
<tr>
<th>Method</th>
<th>Labeled</th>
<th>Un-labeled</th>
<th>Arousal</th>
<th>Valence</th>
<th>Dominance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>KT</td>
<td>Accuracy</td>
<td>KT</td>
</tr>
<tr>
<td>eGeMAPS+2FC</td>
<td>MSP-PODCAST</td>
<td>-</td>
<td>0.372</td>
<td>64.3</td>
<td>0.208</td>
</tr>
<tr>
<td>CBS+2FC</td>
<td>MSP-PODCAST</td>
<td>-</td>
<td>0.417</td>
<td>71.9</td>
<td>0.258</td>
</tr>
<tr>
<td>CBS+LN</td>
<td>MSP-PODCAST</td>
<td>MSP-IMPROV</td>
<td>0.484</td>
<td>79.2</td>
<td>0.313</td>
</tr>
<tr>
<td>CBS+DA</td>
<td>MSP-PODCAST</td>
<td>MSP-IMPROV</td>
<td>0.462</td>
<td>78.5</td>
<td>0.301</td>
</tr>
<tr>
<td>CBS+LN+DA</td>
<td>MSP-PODCAST</td>
<td>MSP-IMPROV</td>
<td>0.506</td>
<td>80.6</td>
<td>0.312</td>
</tr>
<tr>
<td>CBS+LN*</td>
<td>MSP-IMPROV</td>
<td>MSP-IMPROV</td>
<td>0.617</td>
<td>84.7</td>
<td>0.375</td>
</tr>
</tbody>
</table>

**Accuracy: Overall test accuracy**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th>KT</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>eGeMAPS+2FC</td>
<td></td>
<td></td>
<td>0.316</td>
<td>62.7</td>
</tr>
<tr>
<td>CBS+2FC</td>
<td></td>
<td></td>
<td>0.392</td>
<td>69.2</td>
</tr>
<tr>
<td>CBS+LN</td>
<td></td>
<td></td>
<td>0.447</td>
<td>72.3</td>
</tr>
<tr>
<td>CBS+DA</td>
<td></td>
<td></td>
<td>0.442</td>
<td>71.5</td>
</tr>
<tr>
<td>CBS+LN+DA</td>
<td></td>
<td></td>
<td>0.461</td>
<td>74.7</td>
</tr>
<tr>
<td>CBS+LN*</td>
<td></td>
<td></td>
<td>0.558</td>
<td>81.2</td>
</tr>
</tbody>
</table>

### Observations:
- Both adaptation methods lead to improvements
- Best performance achieved by combining ladder network and domain adaptation
Precision as the number of retrieved samples increases

- We evaluate 10% and 20% of the data
- We retrieve samples with low and high values of an attribute
  - Arousal, valence, and dominance
- Success: retrieved samples belong to the correct class created with a median split

Baselines: models using $f()$, trained to predict absolute scores
In this work we explored different preference learning based architectures for SER.
We observed ladder network and Adversarial Domain Adaptation are complementary while adapting SER model to new domain.

This study is supported by Laboratory for Analytic Sciences

Thank You