Towards Conditional Adversarial Training for Predicting Emotions from Speech

Jing Han\(^1\), Zixing Zhang\(^2\), Zhao Ren\(^1\), Fabien Ringeval\(^3\), Björn Schuller\(^{1,2}\)

1. University of Augsburg, Germany
2. Imperial College London, UK
3. Université Grenoble Alpes, France
Motivation

Adversarial training (also called GAN for Generative Adversarial Networks), and the variations that are now being proposed is the most interesting idea in the last 10 years in ML.


Motivation

Towards Conditional Adversarial Training for Predicting Emotions from Speech

**Adversarial Training**

The **Discriminator** tries to distinguish between the generated (fake) and real data.

The **Generator** tries to turn the input noise into fake data to try to fool the discriminator.

\[
\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log (1 - D(G(z)))]
\]
Conditional Adversarial Training

The **Condition** $y$ tries to control the modes of the data being generated, could be based on class labels or data from different modalities, etc.

$$
\min_{G} \max_{D} \mathbb{E}_{x \sim p_{data}(x)} \left[ \log D(x | y) \right] + \mathbb{E}_{z \sim p_{z}(z)} \left[ \log (1 - D(G(z | y))) \right]
$$
Proposed Conditional Adversarial Training for SER

```plaintext
Towards Conditional Adversarial Training for Predicting Emotions from Speech

Proposed Conditional Adversarial Training for SER

Regressors

Discriminators

Loss

Audio features x

Real y

predictions y'

Real

Fake
```
Proposed Conditional Adversarial Training for SER
Towards Conditional Adversarial Training for Predicting Emotions from Speech

Proposed Conditional Adversarial Training for SER

- **Regressor** \((G)\)
- **Discriminator**

**Objective 1:**
\[
\mathbb{E}(\|G(x_t) - y_t\|^2) + \lambda \times \mathbb{E}(\log(D(G(x_t), x_t))) \\
\mathbb{E}(\log(D(y_t, x_t))) + \mathbb{E}(\log(1 - D(G(x_t), x_t)))
\]

**Objective 2:**
(wasserstein distance)
\[
\mathbb{E}(\|G(x_t) - y_t\|^2) + \lambda \times \mathbb{E}[D(G(x_t))] \\
\mathbb{E}[D(x_t)] - \mathbb{E}[D(G(x_t))]
\]
Database

- recordings of spontaneous collaborative and affective interactions in French
- partially used in AVEC 2015 and 2016
- 46 subjects (23 pairs) and 5 minutes per pair
- 16 training / 15 development / 15 test
- http://diuf.unifr.ch/diva/recola/index.html
Experiments

● feature set: mean & var of MFCC 0-12 and log energy

● structure for both NN1 and NN2: LSTM-RNN with 2 layers, 20 nodes per layer.

● measurement: CCC (concordance correlation coefficient)

\[
\rho_{\text{ccc}} = \frac{2 \rho_{\text{pcc}} \sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}
\]
### Experimental Results

<table>
<thead>
<tr>
<th>Model</th>
<th>CCC (dev)</th>
<th>CCC (test)</th>
<th>Arousal (dev)</th>
<th>Arousal (test)</th>
<th>Valence (dev)</th>
<th>Valence (test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNN (2 layers)</td>
<td>.777</td>
<td>.718</td>
<td>.491</td>
<td>.435</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RNN (4 layers)</td>
<td>.761</td>
<td>.723</td>
<td>.487</td>
<td>.390</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conditional adversarial training</td>
<td>.780</td>
<td>.732</td>
<td>.501</td>
<td>.455</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conditional adversarial training with Wasserstein</td>
<td>.797</td>
<td>.737</td>
<td>.474</td>
<td>.444</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Experimental Results

<table>
<thead>
<tr>
<th>Method</th>
<th>CCC</th>
<th>Arousal</th>
<th>Valence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>dev</td>
<td>test</td>
<td>dev</td>
</tr>
<tr>
<td>CCC-objective</td>
<td>.412</td>
<td>.350</td>
<td>.242</td>
</tr>
<tr>
<td>end-to-end</td>
<td>.752</td>
<td>.699</td>
<td>.406</td>
</tr>
<tr>
<td>reconstruction-error based</td>
<td>.785</td>
<td>.729</td>
<td>.378</td>
</tr>
<tr>
<td>prediction-based</td>
<td>.774</td>
<td>.744</td>
<td>.412</td>
</tr>
<tr>
<td>conditional adversarial training</td>
<td>.780</td>
<td>.732</td>
<td><strong>.501</strong></td>
</tr>
<tr>
<td>conditional adversarial training with wasserstein</td>
<td><strong>.797</strong></td>
<td>.737</td>
<td>.474</td>
</tr>
</tbody>
</table>
Experimental Results

(a) arousal

(b) valence

value

-0.5

0.0

0.5

gold standard
proposed method

0 50 100 150 200 250 300
time (s)
Conclusion & Perspective

Conclusion

- adversarial training for speech emotion recognition
- two variants both perform better than baseline
- arousal prediction benefits more from Wasserstein distance than valence

Future work

- train two models based on a loss threshold rather than fixed schedule
- experiments on other dataset
- perform an end-to-end structure
Thank you for listening

jing.han@informatik.uni-augsburg.de