
Wei-Cheng Lin, Kusha Sridhar and Carlos Busso
Outline

1. Background

2. Proposed Methodology

3. Experimental Results and Analysis

4. Conclusions & Future Works
SSL in SER

- Semi-Supervised Learning (SSL)
  - Leverage large amounts of unlabeled data to improve recognition generalization ability
- SSL in Speech Emotion Recognition (SER)
  - Reconstruction-based architecture (i.e., AE [1], VAE [2] and LadderNet [3])

Uninterpretable Hidden Representation of the Bottleneck layer

Typically not used during inference stage (wasteful of training resources)

Emotional Discriminator

ENC

DEC

Unsupervised Path
Proposed Methodology

- DeepEmoCluster framework
  - Inductive SSL scheme (i.e., pseudo-labeling by K-means clustering)

1. No **Decoder** in the framework

2. Meaningful and interpretable hidden representations, resulting in **emotional clusters**!
**DeepEmoCluster**

- **End-to-End SSL framework**
  - Input with 128-Mel spectrogram (32ms window size and 16ms overlaps)
  - **Step 1:** chunk-segmentation pre-processing [Lin and Busso, 2020]

![Diagram showing sentence segmentation and chunking](image)

**Pre-defined Parameters:**
1. $T_{\text{max}}$ (sec): maximum sentence duration in the corpus
2. $w_c$ (sec): desired chunk window length

\[ C = \left\lceil \frac{T_{\text{max}}}{w_c} \right\rceil \]

\[ \Delta c_i = \frac{T_i - w_c}{C - 1} \] (sec): chunk step size depends on sentence duration

No zero-padding is required!

---

DeepEmoCluster

- **Visualization of 128-Mel spectrogram data chunks**
  - Originally arbitrary length of audios are mapped into **fixed size and fixed number of “small spec-images”** as the inputs.
  - These images are shared the same sentence-level emotional target during training procedure.

```
Example of Emotional Target

[“Angry”, “Angry”, …repeat C times]
```

```
0
Δc_i
w_c
T_i (sec)
```

```
C
```
DeepEmoCluster

- **Step 2:** unsupervised (Stage I) + joint-optimization (Stage II)

Some details
1. Stage I is for the unlabeled data
2. Stage II is for the labeled data
3. We reassign the K-means clustering pseudo-labels on every new epoch
Experimental Settings

- Corpus: The MSP-Podcast v1.6
  - Use existing podcast recordings
  - Divide into speaker turns
  - Emotion retrieval to balance the emotional content
  - Annotate using crowdsourcing framework
Experimental Settings

- The MSP-Podcast v1.6
  - 50,362 (83h,29m)
  - Duration range: 2.75 ~ 11 secs
- Corpus partition with minimal speaker overlap sets:
  - Test data: 10,124 samples
    - 50 speakers (25 males, 25 females)
  - Development data: 5,958 samples
    - 40 speakers (20 males, 20 females)
  - Train data: 34,280 samples
    - from remaining speakers
  - Unlabeled data
    - Totally around 500,000 samples
Experimental Settings

- **Parameters Settings:**
  - \( w_c : 1 \) (sec), \( T_{max} : 11 \) (secs) \( \Rightarrow C = 11 \) (sub-images/per sentence)
  - Joint-optimization weighting factor of the loss function \( \lambda = 1 \)
  - 64 batch size, early stopping criteria (for saving the best model based on the min. development loss)
  - \# of K-means cluster = [10, 20, 30] (finetuned parameter depending on the size of unlabeled set)
  - Size of unlabeled set = [0, 15K, 40K] (random sample from the unlabeled data pool)

\[
\text{Loss} = (1 - CCC) + \lambda \times CE
\]
Experimental Results

- Recognition performance under *fully supervised learning* (FSL)
- Baseline Models (all use VGG-16 structure):
  - CNN-regressor
  - CNN-AE (autoencoder)
  - CNN-VAE (variational autoencoder)
- *Statistically significant outperforms baseline models*

<table>
<thead>
<tr>
<th>Model</th>
<th>Aro [CCC]</th>
<th>Dom [CCC]</th>
<th>Val [CCC]</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN-regressor</td>
<td>0.6177</td>
<td>0.4928</td>
<td>0.1696</td>
</tr>
<tr>
<td>CNN-AE</td>
<td>0.6338</td>
<td>0.5111</td>
<td>0.1354</td>
</tr>
<tr>
<td>CNN-VAE</td>
<td>0.5586</td>
<td>0.4800</td>
<td>0.1826</td>
</tr>
<tr>
<td>DeepEmoCluster (10-clusters)</td>
<td><strong>0.6502</strong></td>
<td><strong>0.5426</strong></td>
<td>0.1510</td>
</tr>
</tbody>
</table>
Experimental Results

- Further improved recognition performance while adopting semi-supervised learning (SSL)
  - ADD KEY RESULTS

DeepEmoCluster (10-clusters)

<table>
<thead>
<tr>
<th>Size of Unlabeled Data set</th>
<th>Aro [CCC]</th>
<th>Dom [CCC]</th>
<th>Val [CCC]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 (FSL)</td>
<td>0.6502</td>
<td>0.5426</td>
<td>0.1510</td>
</tr>
<tr>
<td>15K (SSL)</td>
<td>0.6504</td>
<td>0.5400</td>
<td>0.1714</td>
</tr>
<tr>
<td>40K (SSL)</td>
<td>0.6611</td>
<td>0.5400</td>
<td>0.1572</td>
</tr>
</tbody>
</table>
Experimental Results

- Finetuned parameter of # K-means clusters
  - ADD KEY RESULTS

### DeepEmoCluster (SSL-40K)

<table>
<thead>
<tr>
<th># of clusters</th>
<th>Aro [CCC]</th>
<th>Dom [CCC]</th>
<th>Val [CCC]</th>
</tr>
</thead>
<tbody>
<tr>
<td>10-clusters</td>
<td>0.6611</td>
<td>0.5400</td>
<td>0.1572</td>
</tr>
<tr>
<td>20-clusters</td>
<td>0.6491</td>
<td>0.5459</td>
<td>0.1756</td>
</tr>
<tr>
<td>30-clusters</td>
<td>0.6416</td>
<td>0.5490</td>
<td>0.1752</td>
</tr>
</tbody>
</table>
Experimental Results

- Emotional clusters reflecting by the ground-truth emotion distributions under each cluster
  - ADD KEY RESULTS

![Graph showing emotional clusters and distributions](image_url)

**Fig. 3.** Emotional distributions of the clusters with the highest and lowest average level of arousal. The distributions are farther apart with the addition of the supervised SER task.
We introduced a new SSL framework in SER field
DeepEmoCluster achieved the best and competitive recognition performances comparing to other existing SSL frameworks in SER
DeepEmoCluster could result in meaningful hidden representations
We discussed and determined the important parameter
  - The number of K-means clusters is a finetuned parameter depending on the size of unlabeled dataset
Future Works

- Extension of the framework from a single modality (speech) to a multimodal system (speech, language and visual)
  - Forming a comprehensive behavioral emotional clusters

- Strengthen the connections between the latent clusters and the target emotions by utilizing information theory based metric
Release of the MSP-Podcast Corpus

- Academic license
  - Federal Demonstration Partnership (FDP) Data Transfer and Use Agreement
  - Free access to the corpus
- Commercial license
  - Commercial license through UT Dallas

https://msp.utdallas.edu
Reference


Thank you for your attention!

Questions or Contact: wei-cheng.lin@utdallas.edu

This work was funded by NSF (CNS-1823166; IIS-1453781)