MULTI-HEAD ATTENTION FOR SPEECH EMOTION RECOGNITION WITH AUXILIARY LEARNING OF GENDER RECOGNITION

Anish Nediyanchath, Periyasamy Paramasivam, Promod Yenigalla
{anish.n, periyasamy.p, promod.y}@samsung.com
Samsung R&D Institute India - Bangalore, India

ICASSP 2020
Agenda

➢ Motivation & Proposed Method
➢ Raw Audio Processing
➢ Multi Head Attention Network For SER
➢ Multi Head Attention Network For SER + MTL (Gender)
➢ Network Parameters
➢ Dataset
➢ Results Comparison
➢ Conclusion
Motivation & Proposed Method

Motivation
➢ Voice assistants are becoming ubiquitous
➢ Emotional understanding of users makes for better companions
➢ Humans always carries emotion
➢ We express our emotions through
   ➢ What we speak (extrinsic)
   ➢ Voice
   ➢ Expressions (Face, Gesture, Posture etc.)
➢ Targeted for products like chat-bots, voice assistants and social robot

Proposed Method
➢ Transformer Encoder based MHA network for SER with LFBE feature input & Position Embedding
➢ Addition of MTL on MHA network SER with gender prediction as auxiliary task

The effect of emotions on the human voice

<table>
<thead>
<tr>
<th></th>
<th>fear</th>
<th>anger</th>
<th>sorrow</th>
<th>joy</th>
<th>disgust</th>
<th>surprise</th>
</tr>
</thead>
<tbody>
<tr>
<td>speech rate</td>
<td>much faster</td>
<td>slightly faster</td>
<td>slightly slower</td>
<td>faster or slower</td>
<td>very much slower</td>
<td>much faster</td>
</tr>
<tr>
<td>pitch average</td>
<td>very much higher</td>
<td>very much higher</td>
<td>slightly lower</td>
<td>much higher</td>
<td>very much lower</td>
<td>much higher</td>
</tr>
<tr>
<td>pitch range</td>
<td>much wider</td>
<td>much wider</td>
<td>slightly narrower</td>
<td>much wider</td>
<td>slightly wider</td>
<td></td>
</tr>
<tr>
<td>intensity</td>
<td>normal</td>
<td>higher</td>
<td>lower</td>
<td>higher</td>
<td>lower</td>
<td>higher</td>
</tr>
<tr>
<td>voice quality</td>
<td>irregular voicing</td>
<td>breathy chesty tone</td>
<td>resonant</td>
<td>breathy blaring</td>
<td>grumbled chesty tone</td>
<td></td>
</tr>
<tr>
<td>pitch changes</td>
<td>normal</td>
<td>abrupt or stressed stable</td>
<td>downward inflections</td>
<td>smooth upward inflection</td>
<td>voice downward terminal inflection</td>
<td>rising contour</td>
</tr>
<tr>
<td>articulation</td>
<td>precise</td>
<td>tense</td>
<td>slurring</td>
<td>normal</td>
<td>normal</td>
<td></td>
</tr>
</tbody>
</table>
Raw Audio Preprocessing

- 16KHz, 16-bit PCM format audio of 6 seconds input length
- Divide in 46 ms frames with 23 ms stride
- Calculate Log Mel-Filter Bank Energies for each frame, 64 filter banks
- Final feature matrix for training model will be of size 260 * 64
- Scipy wavfile, python speech features library
Multi Head Attention Network For SER

**Attention**

- An **attention** mechanism allows neural network to focus parts on input relevant to given context.
- Attention parameters are a projection of Query (Q) on Key(K)-Value(V) pair vectors.
- Self attention is a variant of attention mechanism where all Q, K, V are from the same input vector.
- Self attention networks replaced sequence based methods like RNN, LSTM and GRU’s.
- We use Multi-Head Attention (MHA), which divides each vector into $n$.
- For audio sequences, neighboring frames will carry similar acoustic characteristics.
- MHA allows model to relate to other parts of the sequence as well if similar characteristics appear.

\[
\begin{align*}
Q_i &= X \ast W_i^Q \\
K_i &= X \ast W_i^K \\
V_i &= X \ast W_i^V \\
H_i &= \text{Softmax} \left( \frac{Q_i K_i^T}{\sqrt{n}} \right) \ast V_i \\
\text{MHA} &= \text{Concat}(H_1, H_2 \ldots H_n) \ast W
\end{align*}
\]
Multi Head Attention Network For SER

**Position Embedding**

- MHA identifies acoustic events relevant for emotion, but not the sequence of events itself
- Position encoding is used in transformer networks to capture sequence information
- Position encoding is fixed positional representation of input features
- BERT extends the idea of position encoding making position information learnable
- We follow the BERT approach by adding a learned position information to input LFBE features
- Position embedding vectors are initialized with random weights and learned as part of training
Multi Task Learning

- MTL allows better generalization by learning to ignore task specific noise patterns
- Contrast between relevant v/s irrelevant features
- Sharing features relevant for different tasks
- Other researches have demonstrated that knowing gender can improve SER
- We introduced gender recognition as auxiliary task to improve SER
- Adding auxiliary MTL task make MHA learn multiple representations relevant in granular space common to both tasks
- Position Embedding and MHA layers are shared by both tasks
- Two independent softmax activations are used for emotion and gender classification
# Network Parameters

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>LFBE Frames</td>
<td>260</td>
</tr>
<tr>
<td>LFBE Features</td>
<td>64</td>
</tr>
<tr>
<td>Input Sequence Length</td>
<td>260</td>
</tr>
<tr>
<td>Batch Size</td>
<td>16</td>
</tr>
<tr>
<td>Number of MHA Layers</td>
<td>6</td>
</tr>
<tr>
<td>MHA Heads</td>
<td>8</td>
</tr>
<tr>
<td>MHA Dropout</td>
<td>0.1</td>
</tr>
<tr>
<td>Feed Forward Layers</td>
<td>6</td>
</tr>
<tr>
<td>Feed Forward Size</td>
<td>256</td>
</tr>
<tr>
<td>MHA Activation</td>
<td>gelu</td>
</tr>
<tr>
<td>Output Activation</td>
<td>softmax</td>
</tr>
<tr>
<td>Learning Rate</td>
<td>0.001</td>
</tr>
<tr>
<td>Optimizer</td>
<td>Adam</td>
</tr>
</tbody>
</table>
Dataset

- IEMOCAP (Interactive Emotional Motion Capture) from University Of Southern California
- It is a multimodal dataset having around 12 hours of audiovisual data
- Consists of five diadic sessions where actors perform improvisations or scriped scenarios to represent emotional expressions
- We use only improvised raw audio samples as it has strong correlation to labelled emotions and are close to natural speech
- Four emotions of Neutral, Happy, Sad and Angry are explored
- IEMOCAP class distributions are skewed – Neutral 49%, Happy 12%, Sad 27%, Angry 12%
- Because of data imbalance, we report both Weighted Average (WA) and Unweighted Average (UA) on test data.
- Dataset is split into 80:20 train and test split for all experiments
- We use five fold cross validation on this dataset for all the reported results
Results from experiments are compared using IEMOCAP dataset and presented.

Comparison is performed against state-of-the-art models using four emotion classes – neutral, angry, happy, and sad.

- MHA attention model increases the overall accuracy by 3% compared to state-of-the-art models.
- Adding position embedding improved results over MHA model by 0.6%, average class accuracy by 2%.
- MTL network with MHA and PE gives overall accuracy of 76.4, 5.3% higher than state-of-the-art by 5.3%.
- The average class accuracy is 70.1, and improvement of 6.2%.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Overall Accuracy (WA)</th>
<th>Class Accuracy (UA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lee [18] (Bi-LSTM)</td>
<td>62.8</td>
<td>63.9</td>
</tr>
<tr>
<td>Satt [19] (CNN + LSTM)</td>
<td>68.8</td>
<td>59.4</td>
</tr>
<tr>
<td>Ramet [20] (Attn. Bi-LSTM)</td>
<td>68.8</td>
<td>63.7</td>
</tr>
<tr>
<td>Zhang [21] (Attn. CNN)</td>
<td>70.4</td>
<td>63.9</td>
</tr>
<tr>
<td>Yenigalla [22] (CNN)</td>
<td>71.5</td>
<td>61.9</td>
</tr>
<tr>
<td>MHA</td>
<td>74.1</td>
<td>64.2</td>
</tr>
<tr>
<td>MHA + PE</td>
<td>74.7</td>
<td>66.2</td>
</tr>
<tr>
<td>MHA + PE + MTL</td>
<td>76.4</td>
<td>70.1</td>
</tr>
</tbody>
</table>
Conclusion

➢ This work demonstrated MHA network for SER
➢ We also employed Multi Task Learning with auxiliary task of gender recognition
➢ Using self attention to attend to different sections of speech features can improve SER accuracy
➢ Demonstrated an overall improvement of accuracy to 76.4%
➢ We believe our work as a step forward for SER towards building conversation systems with better emotion recognition capabilities
➢ In future, we plan to extend this work by bringing in speaker identification to MTL, add more emotions and explore noisy speech
Thank You