Modeling Uncertainty in Predicting Emotional Attributes from Spontaneous Speech

Kusha Sridhar And Carlos Busso
Speech emotion recognition is a hard problem
  ▪ Prediction are not always reliable

Wide application domains – accurate and confident recognition using DNNs

The key challenge is to define mechanisms to quantify reliability to accept or reject an instance
  ▪ e.g., Softmax Response
  ▪ Monte Carlo Dropout

First study on reject option using uncertainty modeling for speech emotion recognition
Ambiguous emotional content leads to low SER performance

It’s important to know what the model does not know
  - Abstain from predicting when in doubt, reducing the risk of error
  - Involve human-in-the-loop

SER models should provide a prediction score along with its confidence
  - We can use confidence to achieve a low error rate while still maintaining coverage as high as possible (reject option)
  - Reliable SER models can be helpful in mission critical applications in (e.g., healthcare and security)
  - Uncertainty prediction facilitates human-in-the-loop solutions
Related Work

**Speech and Image Tasks**
- Selective guaranteed risk algorithm for Imagenet and CIFAR-10 classification tasks [Geifman et. al. 2017]
- Capturing uncertainty from text transcriptions and word error rates to solve ASR task [Dey et. al. 2019, Vyas et.al. 2019]

**Speech Emotion Recognition**
- Use human labelers’ agreement to build emotion scoring models [Deng et. al. 2012]
- Include samples from target domain in a semi-supervised fashion based on confidence levels achieved from multi-corpora training [Deng et. al. 2012]
- Applying reject option to emotion classification under a risk minimization framework: learning thresholds based on softmax response and difference between two highest predictions [Sridhar and Busso 2018]
- Use MC dropout as a sampling technique for active learning to train autoencoder with unlabeled data selected based on their posterior probability estimates [Abdelwahab and Busso 2019]
Outline

1. Dropout for Confidence Estimation
2. Database
3. Analysis
4. Application for Reject Options
DNNs with dropout regularization can be used to quantify prediction uncertainty [Gal et al., 2016]

- We can represent the models’ uncertainty
- Use different configurations of dropout, analyzing predictions per sample
- We can estimate the posterior distribution on the predictions during inferences by sampling weights in a Monte Carlo fashion
Dropout can approximate a Bayesian Inference in deep Gaussian processes [Gal et al., 2016]

- Change the weights setup randomly by applying dropout
- As such, different configurations of the network lead to slightly different prediction
- Prediction Uncertainty will be the variance of N step predictions
- Multiple iterations through a network with dropout is analogous to obtaining predictions form an ensemble of thinner networks.

Goal: Learn the confidence of the model in each of its predictions

**Posterior predictive distribution**

\[ p(x_{test}|X) \approx \int p(x_{test}|\omega)p(\omega|X)d\omega \]

Sample ordered binwise according to uncertainty in prediction
Outline

1. Dropout for Confidence Estimation
2. Database
3. Analysis
4. Application for Reject Options
The MSP-Podcast Database

- Use existing podcast recordings
- Divide into speaker turns
- Emotion retrieval to balance the emotional content
- Annotate using crowdsourcing framework

The MSP-Podcast Database

- **MSP-Podcast**
  - Collection of publicly available podcasts (naturalness and the diversity of emotions)
    - Interviews, talk shows, news, discussions, education, storytelling, comedy, science, technology, politics, etc.
  - Creative Commons copyright licenses
  - Single speaker segments, High SNR, no music, no phone quality
  - Developing and optimizing different machine learning framework using existing databases
    - Balance the emotional content
  - Emotional annotation using crowdsourcing platform

---

**Podcast Audio**
- 16kHz, 16b PCM, Mono

**Diarization**
- Duration filter $2.75s < ... < 11s$

**Emotional Annotation**
- Manual screening
- Emotion retrieval
- Remove telephone quality
- Remove segments with music
- SNR filter
MSP-Podcast corpus version 1.4

With emotion labels: 33,262 sentences (56h, 29m)

Primary emotional classes

Arousal

Valence

Dominance
MSP-Podcast database

- Version 1.4 of the **MSP-Podcast** corpus
  - 33,262 (56h,29m)

- Corpus partition with minimal speaker overlap sets:
  - Test data
    - 9,255 samples from 50 speakers (25 males, 25 females)
  - Validation data
    - 4,300 samples from 30 speakers (15 males, 15 females)
  - Train data
    - Remaining 19,707 samples
Acoustic Features

- **Interspeech 2013 Feature set**
  - 65 low level descriptors (LLD)
  - High Level Descriptors (HLDs) are calculated on LLDs resulting in total of 6,373 features
  - HLDs include:
    - Quartile ranges
    - Arithmetic mean
    - Root quadratic mean
    - Moments
    - Mean/std. of rising/ falling slopes

<table>
<thead>
<tr>
<th>4 energy related LLD</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sum of auditory spectrum (loudness)</td>
<td>prosodic</td>
</tr>
<tr>
<td>Sum of RASTA-filtered auditory spectrum</td>
<td>prosodic</td>
</tr>
<tr>
<td>RMS Energy, Zero-Crossing Rate</td>
<td>prosodic</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>55 spectral LLD</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>RASTA-filt. aud. spect. bds. 1–26 (0–8 kHz)</td>
<td>spectral</td>
</tr>
<tr>
<td>MFCC 1–14</td>
<td>cepstral</td>
</tr>
<tr>
<td>Spectral energy 250–650 Hz, 1 k–4 kHz</td>
<td>spectral</td>
</tr>
<tr>
<td>Spectral Roll-Off Pt. 0.25, 0.5, 0.75, 0.9</td>
<td>spectral</td>
</tr>
<tr>
<td>Spectral Flux, Centroid, Entropy, Slope</td>
<td>spectral</td>
</tr>
<tr>
<td>Psychoacoustic Sharpness, Harmonicity</td>
<td>spectral</td>
</tr>
<tr>
<td>Spectral Variance, Skewness, Kurtosis</td>
<td>spectral</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>6 voicing related LLD</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_0$ (SHS &amp; Viterbi smoothing)</td>
<td>prosodic</td>
</tr>
<tr>
<td>Prob. of voicing</td>
<td>voice qual.</td>
</tr>
<tr>
<td>log. HNR, Jitter (local &amp; $\delta$), Shimmer (local)</td>
<td>voice qual.</td>
</tr>
</tbody>
</table>
1. Dropout for Confidence Estimation
2. Database
3. Analysis
4. Application for Reject Options
### Implementation Details

- **Train separate regression model each for arousal, valence and dominance**
  - DNN with 3 dense layers, 512 nodes per layer
  - SDG optimizer with a learning rate of 0.001
  - Cost function: $1 - \text{CCC}$
  - Input: 6,373D feature vector
  - Output: Prediction score for arousal, valence and dominance

- **Activation functions:**
  - Tanh activation at the hidden layers give the best performance across emotional attributes
  - We also compare reject option performance with tanh and ReLU as activation functions.

- **Evaluation metric: CCC**
  \[
  \rho_c = \frac{2\rho \sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}
  \]
Prediction uncertainty as a function of emotional attributes:
- Train models each for arousal, valence, dominance with dropout and weight regularization
- Obtain test predictions with corresponding uncertainties for each sample
- Design a scatter plot to visualize uncertainty estimates for each test sample – create uniform bins using prediction scores

Observations:
- More ambiguous emotional content observed among neutral samples (middle samples – high uncertainty)
- Samples with extreme emotional content are predicted more confidently
Performance as a function of uncertainty
- Create five subsets according to uncertainty
  - 0-20%: lower uncertainty
  - 80-100% more uncertainty
- Global Selection
- Balanced Selection

Observations:
- Regression performance decreases as uncertainty increases. Ranges of performance are broader for global selection, creating large performance gaps across sets.
1. Dropout for Confidence Estimation
2. Database
3. Analysis
4. Application for Reject Options
Accepting or rejecting samples based on prediction uncertainty
- Rejecting ambiguous samples improves prediction performance of the model but at the same time reduces test coverage

Experiment:
- DNN performance optimized on the validation set with a fixed dropout of 0.5 for all emotional attributes. Here dropout *is not used* during inference.
- Accept or reject a test sample based on prediction uncertainty achieved from MC dropout models. Here dropout *was used* during inference

Performance reported with tanh and ReLU activations at the hidden layers
- **Baseline:** CCC at 100% test coverage without MC dropout

- **Observations**
  - CCC improves as more uncertain samples are rejected, leading to decrease in coverage
  - Reject Option leads to gains in CCC across emotional attributes without compromising too much on coverage
  - Rejecting samples without attempting to balance their emotional content is better
MC dropout is an effective method to quantify uncertainty in SER systems

Confidence of SER models is higher for samples with extreme emotional values

Rejecting samples with low confidence/high uncertainty increases the regression performance

At a test coverage of 75%, relative gains in CCC was observed up to:
- 7.34% (arousal); 13.73% (valence); 8.79% (dominance)

Future Work
- Understanding the impact of different activation functions
- Uncertainty modeling in semi-supervised and unsupervised cases
Thank you

- This work was funded by NSF CAREER Grant IIS-1453781

Our Research: msp.utdallas.edu