Inferring Emotion from Large-scale Internet Voice Data: A Semi-supervised Curriculum Augmentation based Deep Learning Approach

Presentor: Suping Zhou

Suping Zhou, Jia Jia, Zhiyong Wu, Zhihan Yang, Yanfeng Wang, Wei Chen, Fanbo Meng, Shuo Huang, Jialie Shen, Xiaochuan Wang

Department of Computer Science and Technology
Tsinghua University, China
Index

- Introduction
- Related works
- Ideas & contributions
- Methodology
- Experiments
- Conclusion
1 Introduction
Inferring Emotion from Large-scale Internet Voice Data

Introduction

• Internet voice data
  • Internet voice data are users’ speech queries from the Voice Dialogue Applications (VDAs), such as Siri, Sogou Voice Assistant
  • tremendous amounts of VDA users bring in diverse emotion expressions
  • weak and unbalanced emotion expressions.

• acted voice data
  • Traditionally, researches on speech emotion recognition are based on acted voice datasets, which have limited speakers but strong and clear emotion expressions.
Inferring Emotion from Large-scale Internet Voice Data

Introduction

- Inspired by the contrast between the large scale internet voice data from VDAs and the acted emotional voice dataset, can we use one’s strengths to compensate for the other’s weakness?
2 Related works
• Inferring voice emotion
  • In terms of emotion analysis for voice, previous works have focused primarily on extracting effective features and utilizing diverse types of learning methods. (Neumann and Vu 2019) (Freitag et al. 2017) (Zhang et al. 2019)
  • However, all these researches mainly focused on inferring emotions from acted corpora data, few have been done to address the problem for real-world large-scale internet voice data with weak emotion expressions and tremendous uncertain speakers.
  • It is potential to transfer the emphasis on emotion recognition in the wild and assist this work through the augmentation of acted corpus.
Curriculum learning

Related works

- Curriculum learning
  - Curriculum learning is training strategy to learn from simple to complex and proved to achieve great improvements in generalization and speed of convergence. (Bengio et al. 2009)

- It is natural to apply curriculum learning to emotion recognition since we learn to perceive emotions gradually from infancy to adulthood.

- Previous work on speech emotion recognition has utilized curriculum learning to solve the problem of Crowd-sourced Labels and achieve improvements (Lotfian and Busso 2019).
Semi-supervised learning

Related works

• Semi-supervised learning
  • Autoencoders have always been a common way to make better use of unlabeled data in speech emotion recognition. (Deng et al. 2017) (Jia et al. 2018). Some Generative and Adversarial Networks (Semi-VAE (Zhou et al. 2018b), DCGAN (Chang and Scherer 2017), ADDoG (Gideon, McInnis, and Provost 2019) are also utilized to make improvements.

• (Berthelot et al. 2019) propose a hybrid method named Mix-match which combines several ideas and components from the current dominant paradigms for SSL.
3 Ideas & contributions
Challenges

1) how to effectively leverage acted voice dataset with strong and clear emotion expressions to enhance internet voice data?

2) how to utilize large-scale unlabeled data with diverse user emotion expressions to augment few labeled data.
We proposed A Semi-supervised Curriculum Augmentation based Deep Learning Approach

1、Curriculum learning based epoch-wise training strategy

2、Multi-path Mix-match Multimodal Deep Neural Network (MMMD)
4 Methodology
Workflow of our framework

Figure 2: The workflow of our framework.
Supervised MMD

Multi-path Multimodal Deep Neural Network (MMD)

- Multi-path solution
- Multimodal Compact Bilinear pooling
- to model the complex intra-modality relationship which balances both the independencies and dependencies of multi-modal features.
Semi-supervised MMMD

Multi-path Mix-match Multimodal Deep Neural Network (MMMD)

1. Augmentation
   Gaussian noise

2. Mixup
   \[ \lambda \sim \text{Beta}(0.75, 0.75) \]
   \[ \lambda' = \max(\lambda, 1 - \lambda) \]
   \[ x' = \lambda' x_1 + (1 - \lambda') x_2 \]
   \[ p' = \lambda' p_1 + (1 - \lambda') p_2 \]

3. Entropy minimization
   \[ \text{Sharpen}(p, T) = p^\frac{1}{T} / \sum_{j=1}^{L} p_j^\frac{1}{T} \]

MixMatch (Berthelot et al. 2019)
Epoch-wise Training Strategy

\[ X' = \text{Augment}(X) \]
\[ V_t^{U'} = \text{Augment}(V_t^U) \]
\[ W = \text{Shuffle}(\text{Concatenate}(X', V_t^{U'})) \]
\[ X'' = \text{MixUp}(X', W_1) \]
\[ V_t^{U''} = \text{MixUp}(V_t^{U'}, W_2) \]

\[ \mathcal{L}_{V_e} = \frac{1}{|V_e|} \sum_{x, p \in V_e} H(p, P(y|x; \theta)) \]
\[ \mathcal{L}_{V_t} = \frac{1}{|V_t^L|} \sum_{x, p \in V_t^L} H(p, P(y|x; \theta)) \]
\[ \mathcal{L}_X = \frac{1}{|X''|} \sum_{x, p \in X''} H(p, P(y|x; \theta)) \]
\[ \mathcal{L}_U = \frac{1}{L|V_t^{U''}|} \sum_{u, q \in V_t^{U''}} ||q - P(y|u; \theta)||_2^2 \]
\[ \mathcal{L} = \mathcal{L}_{V_e} + \mathcal{L}_{V_t} + \mathcal{L}_X + \lambda_U \mathcal{L}_U \]
5 Experiment
• Internet voice Dataset. (Chinese)
  • a corpus of voice data from Sogou Voice Assistant recorded in 2013 (SVAD13)
  • Speech information, speech-to-text information, social attributes
  • 50,000 unlabeled utterances, 2946 manually labeled utterances
  • Emotion category:
    • Neutral: 49.3%, Happiness: 16.5%, Disgust: 11.0%, Boredom: 8.7%, Anger: 9.8% and Sadness: 4.6%.
  • 5-fold cross validation
• Acted voice Dataset. (Chinese)
  • Speech information, speech-to-text information, social attributes
  • Only when two annotators and the volunteer who read the utterance have same opinion about the emotion labeling, the utterance and its label will be adopted.
  • 2397 labeled utterances
  • Emotion category:
    • *Neutral*: 14.0%, Happiness: 23.8%, Disgust: 17.4%, Anger: 22.6% and Sadness: 22.2%. 5- fold cross validation
• Public dataset **IEMOCAP** (English)
  • Speech information and text information
  • Emotion category:

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Happy</th>
<th>Anger</th>
<th>Sad</th>
<th>Neutral</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utterances</td>
<td>1636</td>
<td>1103</td>
<td>1084</td>
<td>1708</td>
<td>5531</td>
</tr>
<tr>
<td>Proportion(%)</td>
<td>29.6</td>
<td>19.9</td>
<td>19.6</td>
<td>30.9</td>
<td>-</td>
</tr>
</tbody>
</table>

• 10-fold leave-one-speaker-out (LOSO) cross-validation
Feature Extraction
Experiment

• Acoustic feature
  • openSMILE toolkit
  • 1,582 statistic acoustic features
  • the INTERSPEECH 2010 Paralinguistic Challenge

---

<table>
<thead>
<tr>
<th>Low level Descriptors (LLDs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCM loudness</td>
</tr>
<tr>
<td>MFCC [0-14]</td>
</tr>
<tr>
<td>log Mel Frequency Band [0-7]</td>
</tr>
<tr>
<td>Line Spectral Pairs (LSP) Frequency [0-7]</td>
</tr>
<tr>
<td>F0 by sub-harmonic summation</td>
</tr>
<tr>
<td>F0 Envelope</td>
</tr>
<tr>
<td>Voicing probability</td>
</tr>
<tr>
<td>Jitter local</td>
</tr>
<tr>
<td>Jitter difference of difference of periods (DDP)</td>
</tr>
<tr>
<td>Shimmer local</td>
</tr>
</tbody>
</table>

**Table 2.** 38-dimensional frame-level acoustic features

<table>
<thead>
<tr>
<th>Statistics Functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position maximum/minimum</td>
</tr>
<tr>
<td>Arithmetic mean, standard deviation</td>
</tr>
<tr>
<td>Linear regression coefficients 1/2</td>
</tr>
<tr>
<td>Linear regression error quadratic/absolute</td>
</tr>
<tr>
<td>Quartile 1/2/3</td>
</tr>
<tr>
<td>Quartile range 2-1/3-2/3-1</td>
</tr>
<tr>
<td>Percentile 1/99</td>
</tr>
<tr>
<td>Percentile range 99-1</td>
</tr>
<tr>
<td>Up-level time 75/99</td>
</tr>
</tbody>
</table>

**Table 3.** 21 kinds of statistics functions applied on LLDs
Textual feature

- Chinese text
  - Thulac Tool for word segmentation
  - word embeddings is learned with word2vec
  - 31.2 million chinese word as the training corpora
- English text
  - public 300 dimensional vectors
  - Trained on 100 billion words From Google News (Mikolov et al. 2013)
- extract 4200-dimensional utterance-level textual features according to the statistic functions (mean, max) over the above LLDs
Feature Extraction

Experiment

Social feature

- we define 7 query topic types {Chat, Consultation, Joke, Entertainment, Operation, Search and Other} as type features and user query locations as the accent features.

Figure 5: (a) The Topics of User Queries. (b) Top 5 User Locations.
Performance of epoch-wise MMMD

<table>
<thead>
<tr>
<th>Method</th>
<th>Neutral</th>
<th>Sadness</th>
<th>Disgust</th>
<th>Anger</th>
<th>Happiness</th>
<th>Boredom</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNN</td>
<td>0.7072</td>
<td>0.3319</td>
<td>0.2198</td>
<td>0.3755</td>
<td>0.4197</td>
<td>0.2064</td>
<td>0.3768</td>
</tr>
<tr>
<td>SAE</td>
<td>0.7045</td>
<td>0.3971</td>
<td>0.2278</td>
<td>0.3746</td>
<td>0.3945</td>
<td>0.206</td>
<td>0.3841</td>
</tr>
<tr>
<td>MixMatch</td>
<td>0.7079</td>
<td>0.3554</td>
<td>0.2446</td>
<td>0.368</td>
<td>0.4407</td>
<td>0.2366</td>
<td>0.3922</td>
</tr>
<tr>
<td>MMMD-w/o-avd</td>
<td>0.7097</td>
<td>0.3663</td>
<td>0.2104</td>
<td>0.4082</td>
<td>0.4608</td>
<td>0.2104</td>
<td>0.3972</td>
</tr>
<tr>
<td>MMMD-w-avd</td>
<td>0.6936</td>
<td>0.3473</td>
<td>0.2496</td>
<td>0.4135</td>
<td>0.4662</td>
<td>0.2496</td>
<td>0.4028</td>
</tr>
<tr>
<td>epoch-wise MMMD</td>
<td>0.6874</td>
<td>0.3976</td>
<td>0.2511</td>
<td>0.4115</td>
<td>0.4618</td>
<td>0.2772</td>
<td><strong>0.4144</strong></td>
</tr>
</tbody>
</table>

**Table 1:** The F1-Measure of inferring emotion in different classification models.

1. **Semi-supervised MMMD:** MMMD improves the F1 by 3.08% comparing to baseline SAE relatively.

2. **Epoch-wise Training Strategy:** The epoch-wise-MMMD with a epoch-wise learning strategy to leverage AVD improves the F1 by 4.12% relatively.
Performance of MMD

Experiment

Comparison to the state-of-art method on public dataset IEMOCAP

<table>
<thead>
<tr>
<th>Method</th>
<th>[ICASSP, 2017]</th>
<th>[ICASSP, 2019a]</th>
<th>[ICASSP, 2019b]</th>
<th>[ACL, 2019]</th>
<th>Our Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNN</td>
<td>58.8</td>
<td>59.54</td>
<td>53.23</td>
<td>59.76</td>
<td>63.7</td>
</tr>
<tr>
<td>MDNN</td>
<td>62.7</td>
<td>61.8</td>
<td>53.43</td>
<td>-</td>
<td>62.2</td>
</tr>
<tr>
<td>AE-ACNN</td>
<td>66.9</td>
<td>-</td>
<td>59.40</td>
<td>-</td>
<td>66.06</td>
</tr>
<tr>
<td>CNN-LSTM</td>
<td>65.8</td>
<td>65.8</td>
<td>59.63</td>
<td>-</td>
<td>66.37</td>
</tr>
<tr>
<td>Attention-GMU</td>
<td>76.7</td>
<td>75.2</td>
<td>65.9</td>
<td>71.69</td>
<td></td>
</tr>
</tbody>
</table>

Comparing the performance ‘feature A+T’, our proposed method outperforms all the state-of-the-art baseline methods. Especially, for the UA of the ‘feature A+T’, +11.1% compared with [ICASSP, 2019b] using CNN-LSTM and +5.3% compared with [ACL, 2019] using Attention-GMU.

Supervised MMD

unweighted accuracy (UA)

weighted accuracy (WA)
Parameter and Data scalability Analysis.

Experiment

(a) Effects of Lamda. (b) Effects of T. (c) Performance with different amount of unlabeled data.
• the textual information can contribute more to the emotion recognition in the real world VDAs.

• utilize all modalities simultaneously can be more effective to infer emotional utterances.
6 Conclusion
Conclusion
Inferring Emotion from Large-scale Internet Voice Data

• We design a curriculum learning based epoch-wise training strategy to effectively utilize the strong and clear emotion from acted corpus to enhance internet voice data.

• We propose a Multi-path Mixmatch multimodal deep learning method (MMMD) to utilize large-scale unlabeled data to augment few labeled data.

• Our approach turns out to be effective in real-world speech emotion inferring, which can provide more intelligent response in real-world VDA applications.
THANKS

2020.12.19