Multimodal Affective Analysis Using Hierarchical Attention Strategy with Word-Level Alignment

Yue Gu*, Kangning Yang, Shiyu Fu, Shuhong Chen, Xinyu Li, Ivan Marsic
Multimedia Image Processing Lab
Electrical and Computer Engineering Department
Rutgers, The State University of New Jersey
Why the affective analysis is necessary?
Progress of Affective Computing

Affective Analysis

Emotion Recognition
- Happy, Excited
- Sadness
- Anger
- Neutral
- Frustration

Sentiment Analysis
- Strong Positive
- Positive
- Neutral
- Negative
- Strong Negative

Speech Signal Processing
- MFCCs
- Prosody
- Vocal Quality

Natural Language Processing
- BoW
- POS
- CNNs, LSTMs

Multi-Modality
Is multi-modality needed?

- Vocal signal prominence

Oh you don’t like that you are west-sider

Neutral or Frustration
Is multi-modality needed?

- Vocal signal prominence

Oh you don’t like that you are west-sider

Happy
Is multi-modality needed?

- Vocal signal prominence

  Oh you don’t like that you are west-sider

  Happy

- Acoustic ambiguity

  “I love this city!”

  “I hate this city!”
Challenges: Feature Extraction

- Gap between features and actual affective states
  - Lack of high-level associations
  - Not all parts contribute equally
Challenges: Modality Fusion

- Decision-level Fusion
  - Lack of mutual association learning

- Feature-level Fusion
  - Fail to learn time-dependent interactions
  - Lack of consistency
Proposed Solutions

- Feature Extraction
  - Hierarchical attention based bidirectional GRUs

- Modality Fusion
  - Word-level fusion with attention
  - An End-to-End multimodal network
Data Pre-processing

- Text Branch
  - Word Embedding: word2vec

- Audio Branch
  - Mel-frequency spectral coefficients (MFSCs)

- Synchronization
  - Word-level forced alignment
Word-level Fusion

\[ f_{\alpha_{ij}} = \frac{\exp(f_{e_{ij}}^T v_f)}{\sum_{k=1}^{L} \exp(f_{e_{ik}}^T v_f)} \]

\[ f_{e_{ij}} = \tanh(W_f f_{h_{ij}} + b_f) \]
Word-level Fusion

(a) Horizontal Fusion

(b) Vertical Fusion

(c) Fine-tuning Attention Fusion

\[ u_\alpha_i = \frac{\exp(u_e_i^T v_u)}{\sum_{k=1}^{N} \exp(u_e_k^T v_u)} + s_\alpha_i \]

- **\( v_i \)**: Dense Layer
- **\( t_{\alpha_i} \)**: Word-level textual attention distribution
- **\( w_{\alpha_i} \)**: Word-level acoustic attention distribution
- **\( t_{h_i} \)**: Word-level textual contextual state
- **\( w_{h_i} \)**: Word-level acoustic contextual state
Baselines

- Sentiment Analysis
  - BL-SVM, LSTM-SVM
  - C-MKL, TFN, LSTM(A)

- Emotion Recognition
  - SVM Trees, GSV-eVector
  - C-MKL, H-DMS

- Fusion
  - Decision-level, Feature-level (utterance-level)
Sentiment Analysis Result

MOSI

![Bar chart showing weighted accuracy and F1 scores for various models including BL-SVM, LSTM-SVM, C-MKL1, TFN, LSTM(A), UL-Fusion*, DL-Fusion*, Ours-HF, Ours-VF, and Ours-FAF. The chart compares weighted accuracy and weighted F1 scores across these models.]
Emotion Recognition Result

IEMOCAP

Weighted Accuracy  Unweighted Accuracy

SVM Trees  GSV-eVector  C-MKL2  H-DMS (5)  UL-Fusion*  DL-Fusion*  Ours-HF  Ours-VF  Ours-FAF  Ours-FAF (5)
Multimodal architecture is needed

**MOSI**

- Weighted Accuracy
- Weighted F1

**IEMOCAP**

- Weighted Accuracy
- Weighted F1
Generalization

MOSI to YouTube

- Ours-HF: Weighted Accuracy 60, Weighted F1 62
- Ours-VF: Weighted Accuracy 64, Weighted F1 66
- Ours-HAF: Weighted Accuracy 68

IEMOCAP to EmotiW

- Ours-HF: Weighted Accuracy 56, Weighted F1 58
- Ours-VF: Weighted Accuracy 59, Weighted F1 61
- Ours-HAF: Weighted Accuracy 62
Attention Visualization

Carry representative information in both text and audio

Successfully combine both textual and acoustic attentions

Label: anger

What about the business what the hell is this

\[ w_{\alpha_i}, t_{\alpha_i}, s_{\alpha_i}, u_{\alpha_i} \]

Word-level acoustic attention distribution
Word-level textual attention distribution
Shared attention distribution
Fine-tuning attention distribution
Attention Visualization

Capture emphasis and importance variation

Vocal signal prominence

Label: happy

Oh you don’t like that you’re west-sider

<table>
<thead>
<tr>
<th>$w_{\alpha_i}$</th>
<th>Word-level acoustic attention distribution</th>
<th>$s_{\alpha_i}$</th>
<th>Shared attention distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_{\alpha_i}$</td>
<td>Word-level textual attention distribution</td>
<td>$u_{\alpha_i}$</td>
<td>Fine-tuning attention distribution</td>
</tr>
</tbody>
</table>
Summary

- A hierarchical attention based multimodal structure
- The word-level fusion strategies
- Word-level attention visualization
Thank you!

Email: yg202@scarletmail.rutgers.edu
Homepage: www.ieyuegu.com