Multimodal Fusion with Attention Mechanism for Trustworthiness Prediction in Car Advertisements

Van Thong Huynh  
Chonnam National University  
Gwangju, Korea  
207081@jnu.ac.kr

Hyung-Jeong Yang  
Chonnam National University  
Gwangju, Korea  
hjyang@jnu.ac.kr

Guee-Sang Lee  
Chonnam National University  
Gwangju, Korea  
gslee@jnu.ac.kr

Jung-Hee Kim  
North Carolina A&T State University  
North Carolina, United State  
jungkim@ncat.edu

Soo-Hyung Kim  
Chonnam National University  
Gwangju, Korea  
shkim@jnu.ac.kr

ABSTRACT
In this paper, we present our approach to estimate the trustworthiness intensity, a kind of affective state, in advertisements. Our method explored multi-modal (audio, video, text) fusion with LSTM to learn the relationship between frames in video, and attention mechanism to fuse the learned representation of these features. We achieved a CCC score of 0.3426 on validation set of MuSe-Car dataset which outperform baseline methods. In terms of test set, we reached a promising result of 0.3353.

CCS CONCEPTS
- Computing methodologies → Activity recognition and understanding.

KEYWORDS
trustworthiness prediction, affective computing, facial expression, multi-modal

1 INTRODUCTION
Trustworthiness play an important role in the communication between people in daily talks, discussion, or advertisement. They are the key component to help us make decision, but it is not to easy to estimate the reliability in the real world because it depends on many factors as well as subjective or objective reasons. Hence, the development of an automatic system which ability to estimate the trustworthiness is necessary to eliminate the bias in judgments.

Over the past few decades, there are not much research about this aspect. Águado et al. [1] performed two studies about the relationship between trustworhiness and continuous emotion of face. They reported that valence is positive while arousal is negative correlated to trustworthiness judgments. Stappen et al. [12] deployed early fusion LSTM-RNN with self-attention to tackle this problem with features of audio signal, language, and images. It included two parallel LSTM-RNNs to encode the two corresponding query and value vector sequences.

In this study, we present an system in the end-to-end learning to estimate the trustworthiness in frame-level of videos which focused on the car advertisement. In the rest of the paper, we describe the dataset in section 2, our approach in section 3. The results and summary are present in section 4 and 5.

2 DATASET
Our experiments are conducted on MuSe-Car dataset [12], a large and multimodal dataset for understanding multimodal sentiment analysis in-the-wild. It contains videos of vehicle reviews from professional, semi-professional, and casual reviewers in the middle of 20s to the late 50s. Each video is annotated with a continuous value in range of [−1, 1] for trustworthiness in frame-level by at least 5 independent annotators. The details of data distribution in MuSe-Car dataset for trustworthiness prediction are shown in Table 1.

<table>
<thead>
<tr>
<th>Partition</th>
<th>Num of videos</th>
<th>Total length</th>
<th>Num of samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>166</td>
<td>22h 45m 52s</td>
<td>340994</td>
</tr>
<tr>
<td>Devel</td>
<td>62</td>
<td>06h 32m 22s</td>
<td>99155</td>
</tr>
<tr>
<td>Test</td>
<td>64</td>
<td>06h 12m 53s</td>
<td>89686</td>
</tr>
<tr>
<td>Total</td>
<td>292</td>
<td>35h 51m 07s</td>
<td>529835</td>
</tr>
</tbody>
</table>

Table 1: Partitioning of MuSe-Car dataset for trustworthiness prediction.

In together with raw data (visual, audio, text), [12] also provided 8 feature sets extracted from these data. Acoustic feature sets from audio involve deep learned representation of spectral images from speech instances [2], and eGeMAPS (extended Geneva Minimalistic Acoustic Parameter Set) [7] from OpenSMILE toolkit [8]. The visual feature sets contain 5 kinds: 2D pose key points of a person [5], GoCar [13] - the localisation of 28 car parts, environment features from deep network - Xception [6], facial features (landmarks, head pose, action units) from OpenFace [3], and deep representation of facial by pre-trained of VGG16 [11] on VGGFace dataset [10]. Text features are obtained with FastText [4] - a library for efficient learning of word embeddings.
3 NETWORK ARCHITECTURE

In this work, we examined feature sets in an end-to-end multi-modal fusion. We used the same unimodal for each kind features as in Figure 1. We used 3 LSTM layers on each segment of length $m$ to capture the temporal relationship between samples (frame) from that segment in each video, and 2 fully connected (FC) layers as regression layers based on knowledge encoded by LSTM layers. We used attention mechanism to determine the contribution of each unimodal in the fusion step before the last FC layer which the trustworthiness intensity for each sample. The fusion process can be formulated as

$$\mathcal{F}(X_1, X_2, \ldots, X_k) = \sum_{i=1}^{k} W_{att_i} X_i, \quad (1)$$

where $X_i$ is the output of unimodal $i^{th}$. The contribution $W_{att_i}$ are learnable by the following equation

$$W_{att} = \sigma(W_2 \tanh(W_1 X)), \quad (2)$$

with $X$ is the concatenate of $X_i$ and has size of $m \times D \times k$, where $k$ is the number of unimodal, and $D$ is the number of features produce by last FC layer in each unimodal.

4 EXPERIMENTAL RESULTS

The concordance correlation coefficient (CCC) ($\rho_c$) [9] was used to evaluate the performance of our method

$$\rho_c(y, \hat{y}) = \frac{2\sigma(y, \hat{y})}{\sigma(y, y) + \sigma(\hat{y}, \hat{y}) + (\mu - \hat{\mu})^2} \quad (3)$$

where $\mu, \hat{\mu}$ were mean values of $y$ and $\hat{y}$, $\sigma$ denoted the covariance. As the value of $\rho_c$ belong in range of $[-1, 1]$, we used the following function as objective to optimize our network.

$$\mathcal{L}(y, \hat{y}) = 1 - \rho_c(y, \hat{y}) \quad (4)$$

We experimented our method with TensorFlow 2.2 in Python 3.7. We use SGD optimizer with momentum of $1e^{-4}$. The learning rate is gradually increase from 0 to $1e^{-2}$ in first 5 epochs, and decrease to $1e^{-2}$ with cosine scheduler in 36 epochs, and keep as a constant of $1e^{-3}$ from epoch 41$^{st}$ to 50$^{th}$ as in Figure 2.

We conducted experiments with VGGFace, FastText, Deepspectrum, and 2D facial landmarks. We deployed 2 model, in which included one feature set from visual, audio, and text. As the results in Table 2, our model out perform the combination of features with baseline method which proposed in [12] by a large margin except the last row in terms of test data.

5 CONCLUSION

In this paper, we present an approach for trustworthiness intensity estimation with multi-modal features (audio, visual, text) with attention mechanism. Our method achieved the best results on development data, and promising results in test data compare to the baseline methods.

ACKNOWLEDGMENTS

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF2018R1D1A3A03000947, NRF-2020R1A4A1019191). The corresponding author is Soo-Hyung Kim.
Table 2: A summary of our result on MuSe-Trust dataset.

<table>
<thead>
<tr>
<th>Feature set</th>
<th>$\rho_c$ - Validation</th>
<th>$\rho_c$ - Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGGFace + FastText + Deepspectrum</td>
<td>0.3193</td>
<td>0.3353</td>
</tr>
<tr>
<td>Landmarks_2D + FastText + Deepspectrum</td>
<td>0.3426</td>
<td>0.3259</td>
</tr>
<tr>
<td>FastText - Baseline [12]</td>
<td>0.256</td>
<td>0.1343</td>
</tr>
<tr>
<td>eGeMAPS + FastText - Baseline [12]</td>
<td>0.2278</td>
<td>0.2549</td>
</tr>
<tr>
<td>aV (All features from OpenFace) - Baseline [12]</td>
<td>0.1167</td>
<td>0.1378</td>
</tr>
<tr>
<td>eGeMAPS + FastText + aV - Baseline [12]</td>
<td>0.1245</td>
<td>0.1695</td>
</tr>
<tr>
<td>FastText + VGGFace + Raw audio - Baseline [12]</td>
<td>0.3198</td>
<td>0.4128</td>
</tr>
</tbody>
</table>

REFERENCES