SELECT-ADDITIVE LEARNING:
IMPROVING GENERALIZATION IN MULTIMODAL SENTIMENT ANALYSIS

Haohan Wang, Aaksha Meghawat, Louis-Philippe Morency and *Eric P. Xing*

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
{*haohanw, aaksham, morency, epxing*}*@cs.cmu.edu*

ABSTRACT

Multimodal sentiment analysis is drawing an increasing amount of attention these days. It enables mining of opinions in video reviews which are now available aplenty on online platforms. However, multimodal sentiment analysis has only a few high-quality data sets annotated for training machine learning algorithms. These limited resources restrict the generalizability of models, where, for example, the unique characteristics of a few speakers (e.g., wearing glasses) may become a confounding factor for the sentiment classification task. In this paper, we propose a Select-Additive Learning (SAL) procedure that improves the generalizability of trained neural networks for multimodal sentiment analysis. In our experiments, we show that our SAL approach improves prediction accuracy significantly in all three modalities (verbal, acoustic, visual), as well as in their fusion. Our results show that SAL, even when trained on one dataset, achieves good generalization across two new test datasets.

Index Terms— multimodal, sentiment analysis, cross-datasets, generalization, cross-individual

1. INTRODUCTION

Sentiment analysis is the automatic identification of the private state of a human mind with a focus on determining whether this state is positive, negative or neutral [1]. It has been extensively studied in the last few decades [2], primarily based on textual data. With the recent proliferation of online avenues for sharing multimedia content, people are posting more and more videos with opinions. The opinions are expressed through the spoken word (verbal modality), how these words are spoken (acoustic modality) and what gestures and facial expressions accompany the spoken words (visual modality). Multimodal sentiment analysis extends traditional textual sentiment analysis by analyzing all three modalities present in online videos, including acoustic and visual modalities [3, 4].

To foster research in this area, a few datasets have been created with quality annotations for sentiment [1, 5, 6], but

| (a) An illustrative data set | (b) Rules learned |
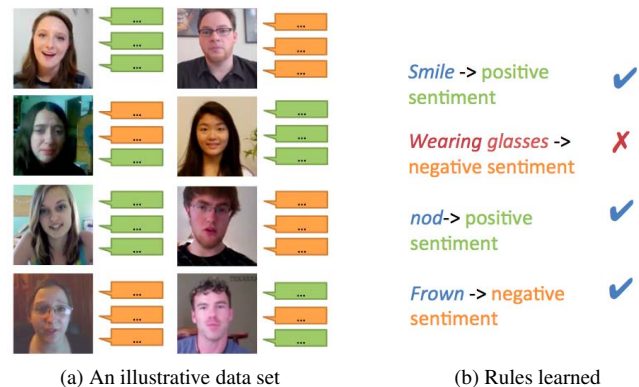
Fig. 1: An illustrative data set demonstrating the "wearing glass" as a confounding factor. Due to the limited amount of data, the model learns that wearing glasses means negative sentiment, which is only applicable to this training data set. (Orange denotes negative sentiment; green denotes positive sentiment; blue denotes correct rules & red denotes incorrect rules).

unfortunately the total number of annotations is still in the order of thousand samples. These limited-size resources make it challenging for conventional machine learning algorithms to generalize well across datasets. In these limited data scenarios, a unique characteristic of a few speakers in the training dataset (e.g., wearing glasses) can end up creating a confounding effect with the sentiment classification task. Fig. 1(a) shows one illustrative example where limited data can bring in learning and generalization challenges. Since in this example all individuals with glasses happen to be expressing negative sentiments, the classifier ends up learning an association between visual appearance of wearing glasses and negative sentiment (see Fig. 1(b)) .

The role that the visual appearance plays here is statistically known as a confounding factor [7, 8]. To generalize across datasets and individuals, a robust multimodal sentiment classifier should not include features from a confounding factor. In other words, the prediction of the sentiment po-

ICME 2017

larity should not be dependent on the speaker's unique characteristic, namely, identity of the speaker.

Before going further in this research agenda, we studied if the confounding factor also exists in the real-world multimodal sentiment analysis datasets. In the MOSI multimodal sentiment analysis data set [6], we tested for the null hypothesis that sentiment is independent of an individual's identity. Chi-square test obtains a p-value of 1.202×10^{-19}, which strongly suggests the dependence between individual identities and the expressed sentiment. Consequently, naively applying machine learning algorithms on this dataset will most likely result in a suboptimal model that misinterprets an individual's identity as prescient information for sentiment.

In this paper, we propose a Select-Additive Learning (SAL) procedure that addresses the confounding factor problem, specifically for neural architectures such as convolutional neural networks. Our proposed SAL approach is a two-phase procedure with the (*Selection* phase and *Addition* phase. During the *Selection* phase, SAL identifies the confounding factors from the latent representation learned by neural networks. During the *Addition* phase, SAL forces the original model to discard (or rather, give less importance to) the confounding elements by adding Gaussian noises to these representations. We conduct extensive experiments to test the performances of state-of-art neural-based models enhanced by SAL. All our experiments are performed in a person-independent setting, where subjects in the test set are different from the training and validation sets. We test the generalization with both, within-data and across-datasets experiments.

2. RELATED WORK

Multimodal data has been studied for a variety of applications to analyze human behaviors, including person detection and identification [9, 10], human action recognition [11, 12], face recognition [13, 14], as well as sentiment analysis.

Originating from analysis of the textual modality, sentiment analysis has been carried out at the word level [15], phrase level [16] and sentence level [17]. [18] performed sentiment analysis on audio data by first transcribing the spoken words and then performing sentiment analysis. Related to audio-based sentiment analysis is the task of estimating emotional state of the speaker from audio input [19]. For the visual modality, the Facial Action Coding System [20] laid the groundwork for analyzing facial expressions and emotions. Recently, convolutional neural networks were used to discover the affective regions for sentiment on still images [21].

The fusion of textual, acoustic and visual modalities for sentiment analysis has drawn increasing attention lately [1]. A variety of methods have been proposed and extensively discussed in recent years [22, 23, 24]. The state-of-the-art performance is achieved with a Convolutional Neural Network [25].
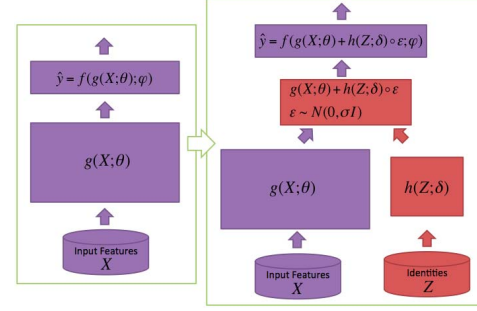
Fig. 2: The SAL architecture is achieved by a simple extension of a general deep learning discriminative classifier. The purple part is the original deep learning model. The red part is the extension SAL introduces. The extension network is connected to the original network via a Gaussian Sampling Layer.

Our proposed Select-Additive Learning (SAL) procedure improves the generalizability of neural networks. Our experiments show improved prediction accuracy for all three modalities (verbal, acoustic and visual) as well as for multimodal fusion. The following section introduces our proposed Select-Additive Learning procedure.

3. SELECT-ADDITIVE LEARNING

The main goal of our work is to increase the generalizability of multimodal sentiment prediction models by encouraging the model to consider sentiment-associated features (i.e. people are smiling while expressing positive sentiment) more than the identity-related features (i.e. wearing glasses).

We formalize the problem by defining an input feature matrix X of size $n \times p$ that encodes the p features for n utterances. In the multimodal scenario, p will be the total number of verbal, acoustic and visual features. We also define a vector y of size $n \times 1$ which represents the sentiment of each utterance. Finally, we define a new matrix Z which encodes for each utterance the speaker identity in a one-hot matrix of size $n \times m$ where m represents the total number of unique individuals in the dataset.

3.1. Select-Additive Learning Architecture

Our proposed SAL procedure is designed to enhance a pre-existing (i.e. pre-trained model) discriminative neural network to be more robust against confounding factors. To formally introduce our SAL approach, we define two main components present in most discriminative neural network classifiers (e.g., Convolutional Neural Network, CNN): a representation learner component and a classification component. To simplify the notation, we use $g(\cdot; \theta)$ denotes the representation learner component and θ stands for its parameters. Our hypothesis is that confounding factors will be constrained to

a subset of dimensions present in $g(\cdot;\theta)$. Similarly, we use $f(\cdot;\phi)$ to denote the classification component and ϕ denotes the parameters. Therefore, a full neural network classifier is denoted as $f(g(\cdot;\theta);\phi)$. In our SAL approach, $g(\cdot;\theta)$ from identity-related features as *identity related confounding dimensions*. Our SAL approach can be summarized as first identifying these dimensions and then reduce the impact of these dimensions by adding noise to them.

To select *identity-related confounding dimensions*, SAL introduces a simple neural network (denoted by $h(\cdot;\delta)$, where δ stands for its parameters). This is to predict *identity-related confounding dimensions* from individual identities Z, by minimizing the difference between $h(Z;\delta)$ and $g(X;\theta)$. Therefore, $h(Z;\delta)$ will effectively pinpoint the *identity-related confounding dimensions* in $g(X;\delta)$. Figure 3a shows an overview of this Selection Phase.

To force the model to discard *identity-related confounding dimensions*, SAL introduces Gaussian noise to these dimensions while minimizing prediction error, so that $f(\cdot;\phi)$ learns to neglect noised representation. The noise is added through a Gaussian Sampling Layer [26]. Figure 3b shows an overview of this addition phase.

Figure 2 shows how SAL assembles $g(\cdot;\theta)$, $f(\cdot;\phi)$ and $h(\cdot;\delta)$ together via a Gaussian Sampling Layer.

3.2. Select-Additive Learning Algorithm

A pre-requisite to our Select-Additive Learning (SAL) approach is first learn a discriminative neural classifier. On our experiments, we achieve this goal by minimizing the following lost function:

$$\operatorname*{argmin}_{\phi,\theta} \frac{1}{2}(y - f(g(X;\theta);\phi))^2$$

The same loss function is often used in discriminative neural networks [27].
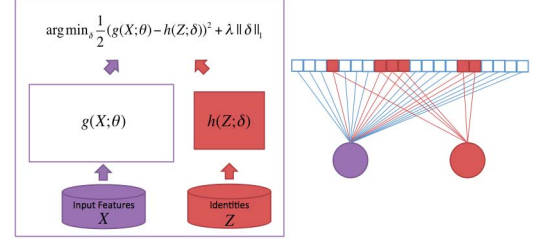
3.2.1. Selection Phase

Once the original representation g(X;θ) is learned, the selection phase optimizes a new loss function to discover the identity related confounding dimensions. This selection phase is operationalized by tuning the parameters λ using the following loss function (as illustrated in Figure 3 (a)):

$$\arg\min_{\delta} \frac{1}{2}(g(X;\theta) - h(Z;\delta))^2 + \lambda||\delta||_1 \qquad (1)$$

where λ is a scalar that controls the weight of the sparsity regularizer. In this phase, both X and Z are available, but only δ is tuned, as shown in Fig. 3(a).

The goal of this phase is to select *identity-related confounding dimensions* from the original representation. To achieve this, we tune δ to minimize the difference between $g(X;\theta)$ and $h(Z;\delta)$. As Z only encodes identity information,

(a) *Selection* Phase: SAL forces $h(Z;\delta)$ to identify *identity-related confounding dimensions*. Right side figure shows that $h(Z;\delta)$ selects these dimensions.
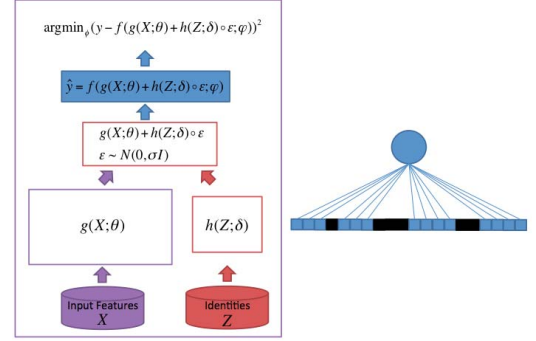
(b) *Addition* Phase: SAL forces the model to focus on other dimensions by adding Gaussian noise to *identity-related confounding dimensions*. Right side shows that the model shifts focus because these dimensions are contaminated/noisy.

Fig. 3: Illustration of SAL. On the left, network structure and training objective is presented. On the right, circles denote neurons. Squares denote dimensions of representation.

the minimum of difference will be achieved when $h(Z;\delta)$ is matched to the *identity-related confounding dimensions* of $g(X;\theta)$. L1 regularization of δ is necessary to avoid overfitting as output dimension of $h(\cdot;\delta)$ is typically significantly higher than input dimension.

The result of this *selection* phase is shown on the right-hand-side of Fig. 3(a). All the weights of original model (purple circle) are active and connected to every dimension while only some weights of $h(\cdot,\delta)$ (red circle) are active and connected to the *identity-related confounding dimensions*

3.2.2. Addition Phase

After the selection phase $h(\cdot,\delta)$ should be pointing at the *identity-related confounding dimensions*. Our remaining step is to learn a new neural network classifier where the confounding dimensions have "masked". We achieve this by adding Gaussian noise. Our addition phase defines the following loss function to achieve this goal:

$$\operatorname*{argmin}_{\phi} \frac{1}{2}(y - f(g(X;\theta) + h(Z;\delta) \circ \epsilon;\phi))^2 \qquad (2)$$

where $\epsilon \sim N(0,\sigma I)$ and $\circ$ stands for element-wise product, , as showed in Fig. 3(b).

In this phase, parameter ϕ is tuned. The input representation of $f(\cdot; \phi)$ consists of the representation learned from $g(X; \theta)$ and the $h(Z; \delta)$-selected *identity-related confounding dimensions* with Gaussian noise added. The noise ensures that *identity-related confounding dimensions* are no longer informative so that $f(\cdot; \phi)$ can be trained to ignore them.

As illustrated on the right side of Fig. 3(b), *identity-related confounding dimensions* are contaminated with addition of noise. Therefore, the model learns to discard these non-informative dimensions, and its weights get optimized to focus on the rest of the dimensions.

4. EXPERIMENTS

In this section, we perform an extensive set of experiments on three different data sets to see whether SAL can help improve the generalizability of a discriminative neural classifier. Generalizability is tested by performing across-dataset experiments where two of the dataset are kept exclusively for testing. All our experiments follow a person independent methodology where none of the subject from the training data are present in the test datasets.

4.1. Models

We compare the following models:

⋆**CNN**: The state-of-the-art seven layer convolutional neural network architecture used previously for multimodal sentiment analysis [25].

⋆**SAL-CNN**: After the state-of-the-art CNN is fully trained, we use SAL to increase its generalizability and predict sentiment. $h(\cdot, \delta)$ is a neural perceptron [27].

4.2. Datasets

We performed our experiment on three multimodal sentiment analysis data sets:

⋆**MOSI**: This dataset consists of 93 videos obtained from YouTube channels. Each video contains the opinions from one unique individual. The dataset has 2199 utterances manually segmented from online videos of movie reviews. Each utterance was also manually annotated for sentiment label [6].

⋆**YouTube**: This dataset consists of 47 opinion videos with 280 utterances with manually annotated sentiment labels [1]. Each video contains the opinions from one unique individual.

⋆**MOUD**: This dataset consists of 498 Spanish opinion utterances from 55 unique individuals [5].

Although, majority of the data originate from YouTube, they differ in recording quality and the processing done after curation. The verbal features in the MOUD dataset need one extra step of translation from Spanish to English. These three datasets are good candidates to evaluate across-dataset generalization.

Table 1: Within data set experiments

		CNN	SAL-CNN
Unimodal	Verbal	0.678	**0.732**
	Acoustic	0.588	**0.618**
	Visual	0.572	**0.636**
Bimodal	Verbal+Acoustic	0.687	**0.725**
	Verbal+Visual	0.706	**0.73**
	Acoustic+Visual	**0.661**	0.621
All Modalities		0.715	**0.73**

4.3. Feature Extraction

We extracted an embedding for each word using a word2vec dictionary pre-trained on a Google News corpus [28]. The text feature of each utterance was formed by concatenating the word embeddings for all the words in the sentence and padding them with the appropriate zeros to have the same dimension. We set the maximum length as 60 and discarded additional words[1]. For YouTube dataset, we extracted the transcripts using the IBM Bluemix's speech2text API[2]. For MOUD dataset, we translated Spanish transcripts into English transcripts. We used openSMILE [29] to extract the low-level audio descriptors for each spoken utterance. These audio descriptors included the Mel-frequency cepstral coefficients, pitch and voice quality. We processed every frame in each video and used the audio-visual synchrony to identify which frames happen during a specific utterance. We used the CLM-Z library [30] for extracting facial characteristic points.

4.4. Experiment Setup

We remove the netural utterances out of the data set. The first 62 individuals in the MOSI data set are selected as training/validation set. There are around 1250 utterances in total. These utterances are shuffled and then 80% are used for training and 20% used for validation. We have three test datasets. 1) MOSI: 546 utterances from the remaining 31 individuals. 2) YouTube: 195 utterances from 47 individuals and 3) MOUD: 450 utterances from 55 individuals. We use MOSI as training set because it is the largest and most recent dataset among all three.

4.5. Experiment Results

4.5.1. Within data set

Table 1 shows the results for **CNN** and **SAL-CNN** tested on the remaining 31 individuals' data of MOSI. The results indicate that SAL could help to increase the generalizability of the trained model.

---

[1] only around 0.5% utterances in our datasets have more than 60 words
[2] https://www.ibm.com/watson/developercloud/speech-to-text.html

Table 2: Across data set experiments

| | Youtube | | MOUD | |
	CNN	SAL-CNN	CNN	SAL-CNN
Verbal	0.605	**0.657**	0.522	**0.569**
Acoustic	0.441	**0.564**	0.455	**0.549**
Visual	0.492	**0.549**	**0.555**	0.548
Ver+Acou	0.642	**0.652**	0.515	**0.574**
Ver+Vis	0.642	**0.667**	0.542	**0.574**
Acou+Vis	0.452	**0.559**	0.533	**0.554**
All	0.611	**0.667**	0.531	**0.574**
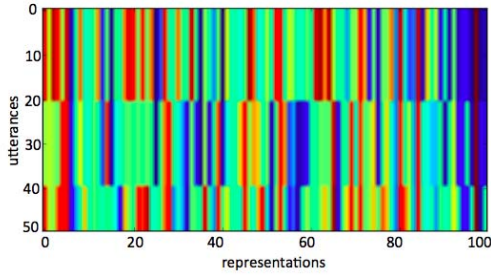
Fig. 4: Confounding factors identified in the *Selection* phase for first 50 utterances (rows), first 100 representation values (columns) in the training set.

4.5.2. Across data sets

Table 2 shows the results for **CNN** and **SAL-CNN** tested on YouTube and MOUD dataset. First, it is noteworthy that in some cases the performance of the CNN is worse than mere chance. This inferior performance substantiates the existence of the non-generalization problems we are targeting.

Overall, Select-Additive Learning increases the robustness and performance of the previous models consistently (except only two cases: Video modality in MOUD and fusion of acoustic & visual in MOSI). Permutation Test rejects the null hypothesis (no improvement) with p-values 0.037, 0.0003, 0.0023 respectively for MOSI, YouTube, and MOUD, indicating significant improvement.[3]

4.6. Discussion

To substantiate our proposed model and algorithm, we examine the learning process and verify that representation of confounding factors exists and our method can mitigate its effects. We demonstrate this with the visual modality as it intuitively contributes the most to confounding.

Figure 4 shows a plot of $h(Z, \delta)$ during the *Selection* phase. It is a zoomed-in figure for the first 50 utterances (rows) and first 100 values of the representation vector (columns). Blue indicates lowest values and red indicates

---

[3]Select-additive Learning implementation is available at https://github.com/HaohanWang/SelectAdditiveLearning

highest values and other colors are linearly interpolated.

The representation of utterances forms clear clusters and each cluster belongs to one person. Despite each individual having their own pattern, some dimensions have generalized well across individuals. Our model learns to assign more weights to these dimensions after noise is introduced.

In addition to these results, we calculated the inter-cluster distance over intra-cluster distance ratio for the representation learned under two situations: 1) clustered by category of sentiment and 2) clustered by individual's identity. We compared the ratios for **CNN** and **SAL-CNN**. The higher ratio indicates a clearer clustering structure.

After SAL, for representation clustered by category of sentiment, the ratio increased by 44%, 15% and 72% respectively for verbal, acoustic, and visual modality, while for representation clustered by individual's identity, the ratio increased by 9%, 3% and 13%, respectively. These numbers indicate SAL almost maintains the clustering structure of identity, but greatly improves the clustering structure of category of sentiment. This shows the effectiveness of SAL.

5. CONCLUSION

High-quality datasets required to train machine learning models for automatic multimodal sentiment analysis are only of the order of a few thousand samples. These limited resources restrict models' generalizability, leading to the issue of confounding factors. We proposed a Select-Additive Learning (SAL) procedure that can mitigate this problem. With extensive experiments, we have shown how SAL improves the generalizability of state-of-the-art models. We increased prediction accuracy significantly in all three modalities (verbal, acoustic, visual), as well as in their fusion. We also showed that SAL could achieve good prediction accuracy even when tested across data sets.

6. REFERENCES

[1] Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *Proceedings of the 13th international conference on multimodal interfaces*. ACM, 2011.

[2] Bo Pang and Lillian Lee, "Opinion mining and sentiment analysis," *Foundations and trends in information retrieval*, 2008.

[3] Akshi Kumar and Mary Sebastian Teeja, "Sentiment analysis: A perspective on its past, present and future," *International Journal of Intelligent Systems and Applications*, 2012.

[4] Martin Wollmer, Felix Weninger, Timo Knaup, Bjorn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe

Morency, "Youtube movie reviews: Sentiment analysis in an audio-visual context," *Intelligent Systems, IEEE*, 2013.

[5] Verónica Pérez Rosas, Rada Mihalcea, and Louis-Philippe Morency, "Multimodal sentiment analysis of spanish online videos," *IEEE Intelligent Systems*, , no. 3, 2013.

[6] Amir Zadeh, "Micro-opinion sentiment intensity analysis and summarization in online videos," in *ICMI*. ACM, 2015.

[7] Robert M Ewers and Raphael K Didham, "Confounding factors in the detection of species responses to habitat fragmentation," *Biological Reviews*, 2006.

[8] Haohan Wang and Jingkang Yang, "Multiple confounders correction with regularized linear mixed effect models, with application in biological processes," in *BIBM*. IEEE, 2016.

[9] Lingxiang Wu, Jinqiao Wang, Guibo Zhu, Min Xu, and Hanqing Lu, "Person re-identification via rich color-gradient feature," in *ICME*. IEEE, 2016.

[10] Xiaoke Zhu, Xiao-Yuan Jing, Fei Wu, Weishi Zheng, Ruimin Hu, Chunxia Xiao, and Chao Liang, "Distance learning by treating negative samples differently and exploiting impostors with symmetric triplet constraint for person re-identification," in *ICME*. IEEE, 2016.

[11] Antonio Tejero-de Pablos, Yuta Nakashima, Tomokazu Sato, and Naokazu Yokoya, "Human action recognition-based video summarization for rgb-d personal sports video," in *ICME*. IEEE, 2016.

[12] Ying Zhao, Huijun Di, Jian Zhang, Yao Lu, and Feng Lv, "Recognizing human actions from low-resolution videos by region-based mixture models," in *ICME*. IEEE, 2016.

[13] Zhongjun Wu and Weihong Deng, "One-shot deep neural network for pose and illumination normalization face recognition," in *ICME*. IEEE, 2016.

[14] Binghui Chen and Weihong Deng, "Weakly-supervised deep self-learning for face recognition," in *ICME*. IEEE, 2016.

[15] Erik Cambria, Daniel Olsher, and Dheeraj Rajagopal, "Senticnet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis," in *AAAI*. AAAI Press, 2014.

[16] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *EMNLP*. Association for Computational Linguistics, 2005.

[17] Ellen Riloff and Janyce Wiebe, "Learning extraction patterns for subjective expressions," in *EMNLP*. Association for Computational Linguistics, 2003.

[18] Lakshmish Kaushik, Abhijeet Sangwan, and John HL Hansen, "Sentiment extraction from natural audio streams," in *ICASSP*. IEEE, 2013.

[19] Boya Wu, Jia Jia, Tao He, Juan Du, Xiaoyuan Yi, and Yishuang Ning, "Inferring usersemotions for human-mobile voice dialogue applications," .

[20] Paul Ekman and Wallace V Friesen, "Facial action coding system," 1977.

[21] Ming Sun, Jufeng Yang, Kai Wang, and Hui Shen, "Discovering affective regions in deep convolutional neural networks for visual sentiment prediction," in *ICME*. IEEE, 2016.

[22] Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency, "Utterance-level multimodal sentiment analysis.," in *ACL*, 2013.

[23] Luca Casaburi, Francesco Colace, Massimo De Santo, and Luca Greco, "magic mirror in my hand, what is the sentiment in the lens?: An action unit based approach for mining sentiments from multimedia contents," *Journal of Visual Languages & Computing*.

[24] Soujanya Poria, Erik Cambria, Newton Howard, Guang-Bin Huang, and Amir Hussain, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content," *Neurocomputing*.

[25] Soujanya Poria, Erik Cambria, and Alexander Gelbukh, "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis," in *EMNLP*, 2015, pp. 2539–2544.

[26] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[27] Haohan Wang and Bhiksha Raj, "On the origin of deep learning," *arXiv preprint arXiv:1702.07800*, 2017.

[28] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[29] Florian Eyben, Martin Wöllmer, and Björn Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *International conference on Multimedia*. ACM, 2010.

[30] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency, "3d constrained local model for rigid and non-rigid facial tracking," in *CVPR*. IEEE, 2012, pp. 2610–2617.